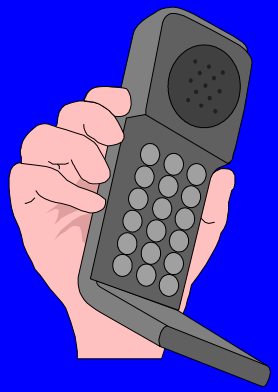
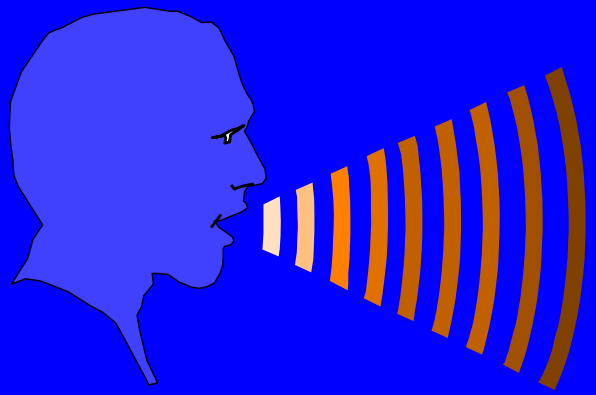
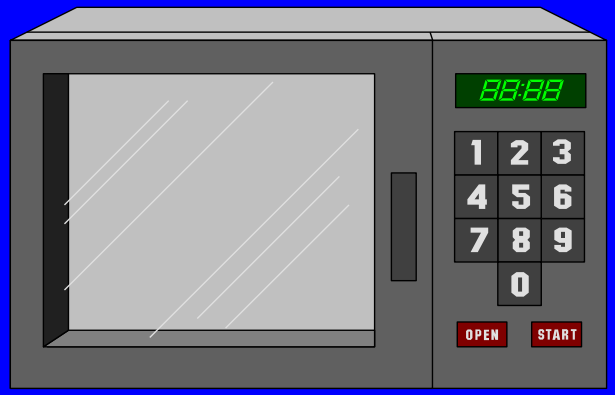
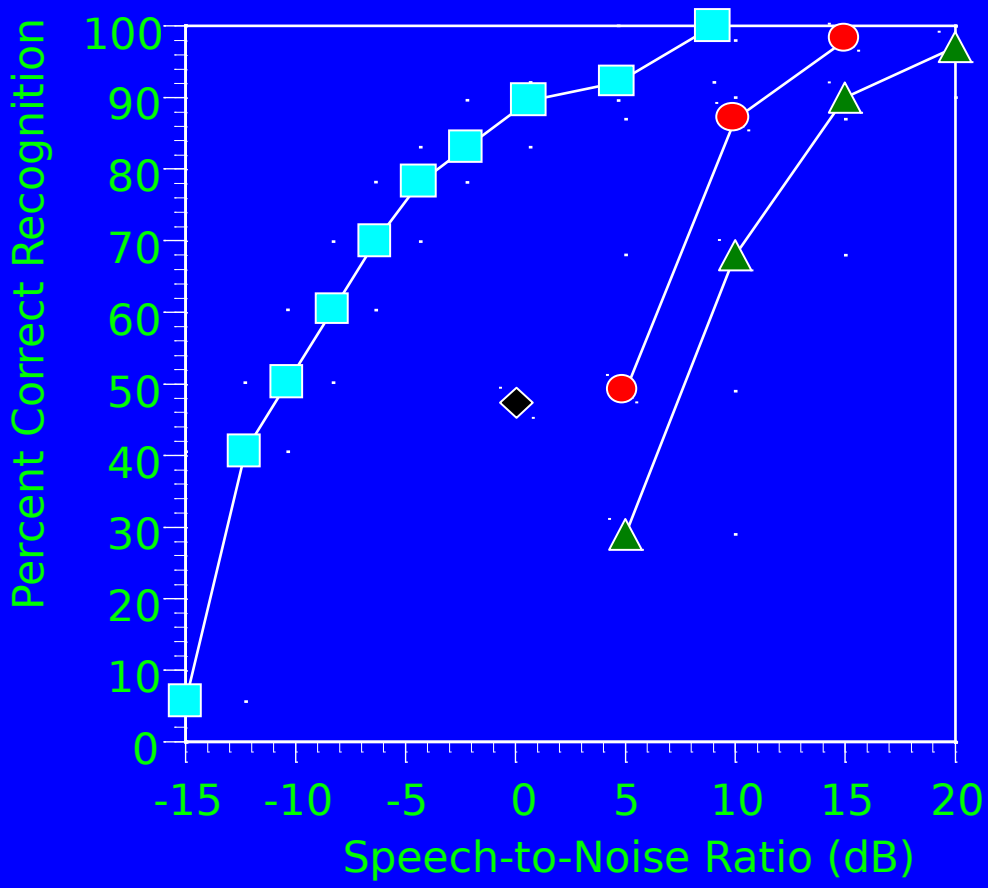
Auditory-Visual Integration: Implications for the New Millennium



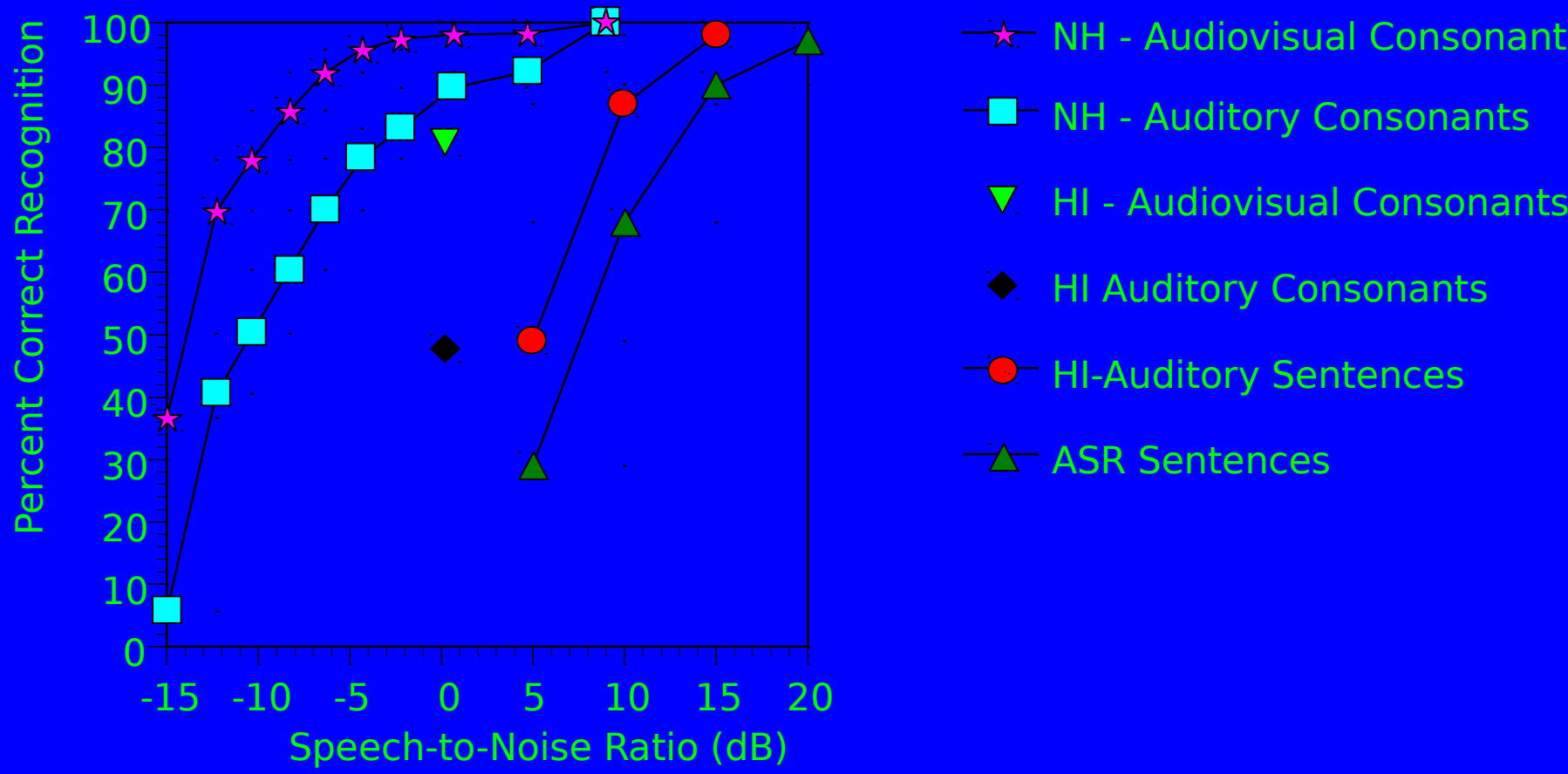




Speech Recognition by Man and Machine



Auditory-Visual vs. Audio Speech Recogniti

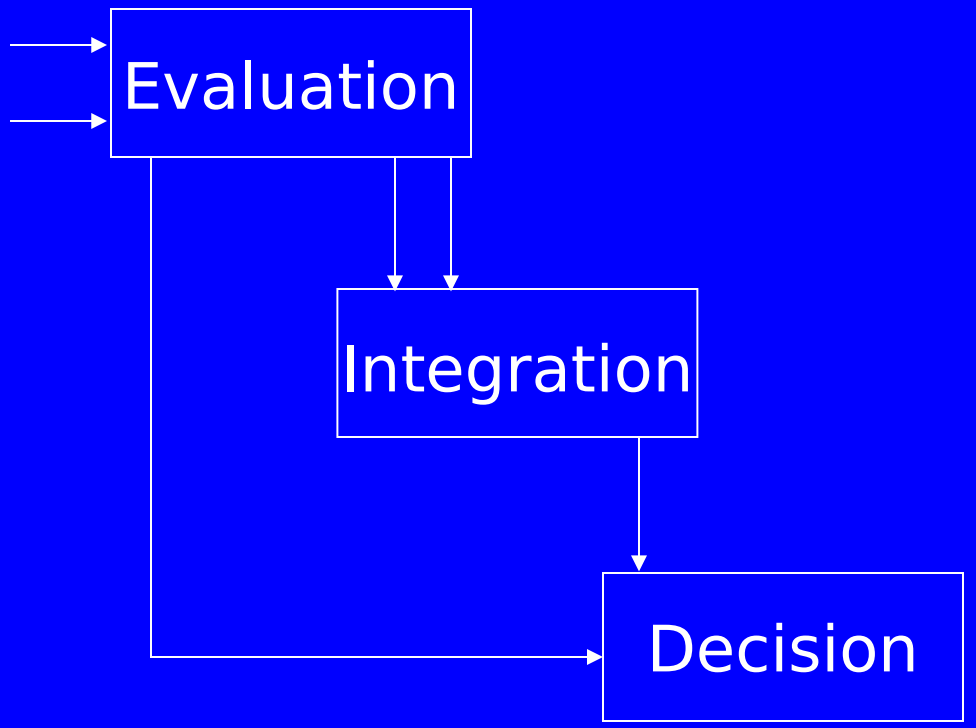


Common Problems – Common Solutions

- Automatic Speech Recognition
- Hearing Aid Technology
- Synthetic Speech
- Learning and Education - Classrooms
- Foreign Language Training

How It All Works - The Prevailing View

- *Information extracted from both sources independently*
- *Integration of extracted information*
- *Decision statistic*



From Massaro, 1998

Possible Roles of Speechreading

- ***Provide segmental information that is redundant with acoustic information***
  - ***vowels***

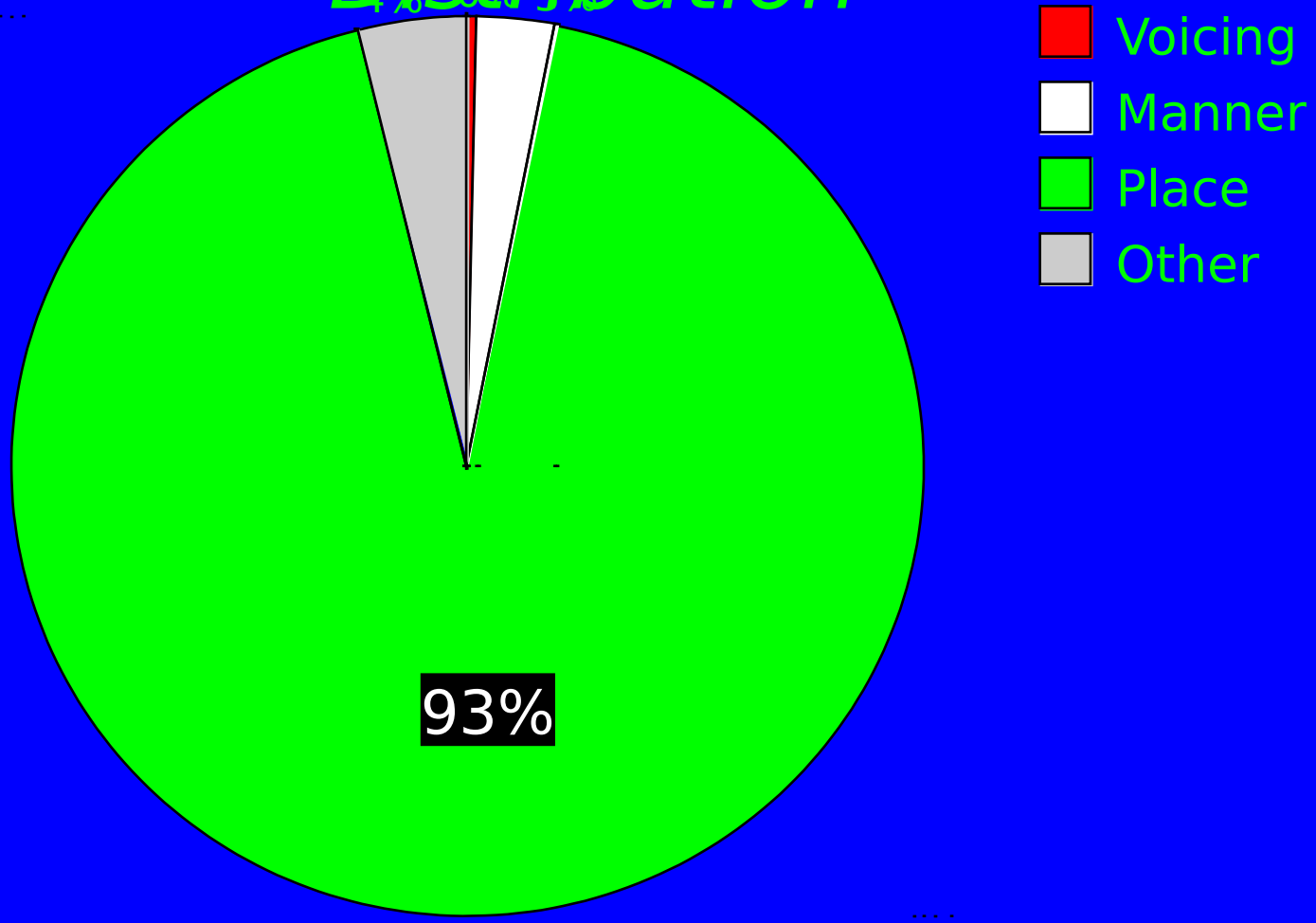
Possible Roles of Speechreading

- Provide segmental information that is redundant with acoustic information
  - vowels
- ***Provide segmental information that is complementary with acoustic information***
  - ***consonants***

Possible Roles of Speechreading

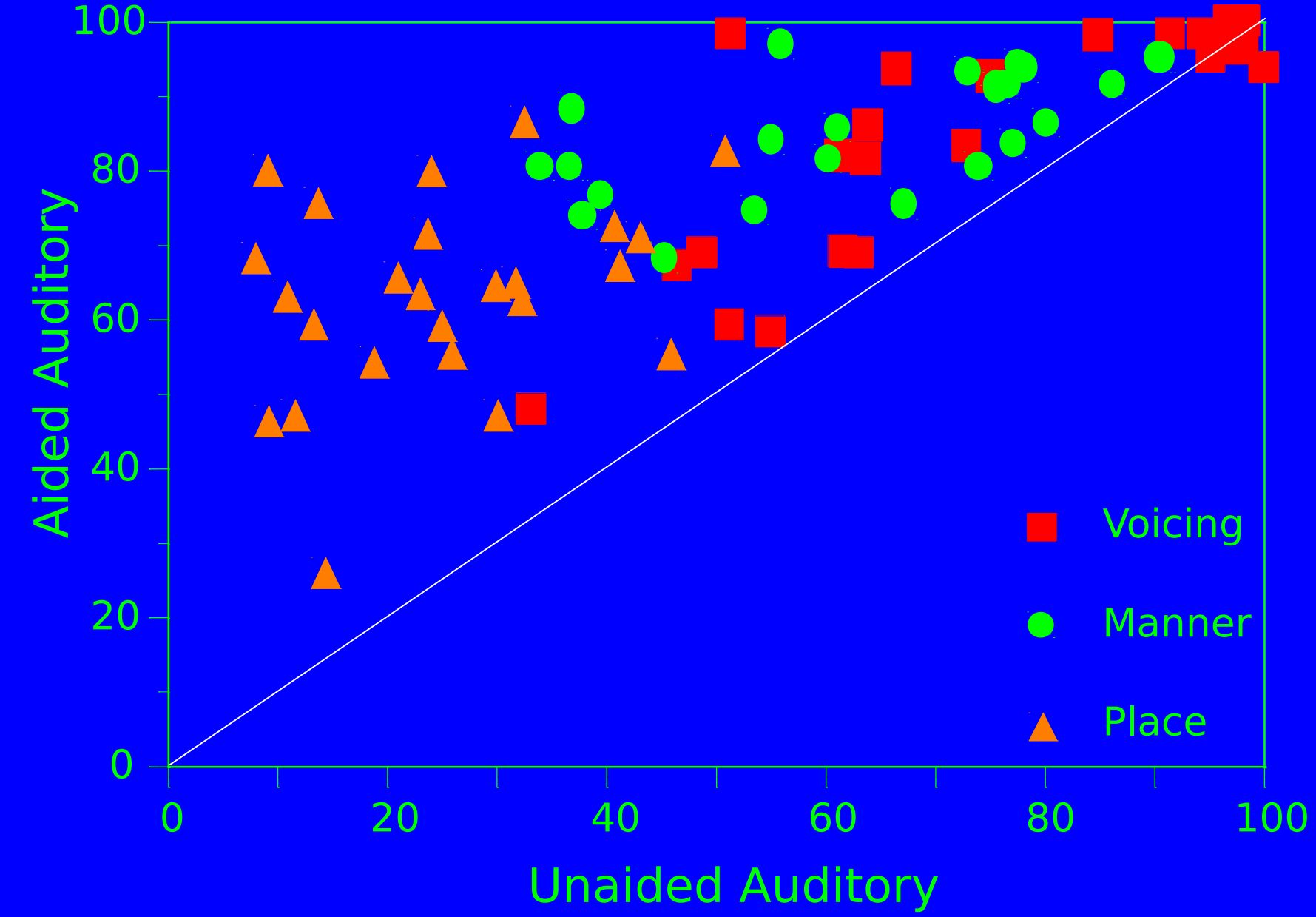
- Provide segmental information that is redundant with acoustic information
  - vowels
- Provide segmental information that is complementary with acoustic information
  - consonants
- ***Direct auditory analyses to the target signal***
  - ***who, where, when, what (spectral)***

Visual Feature Distribution

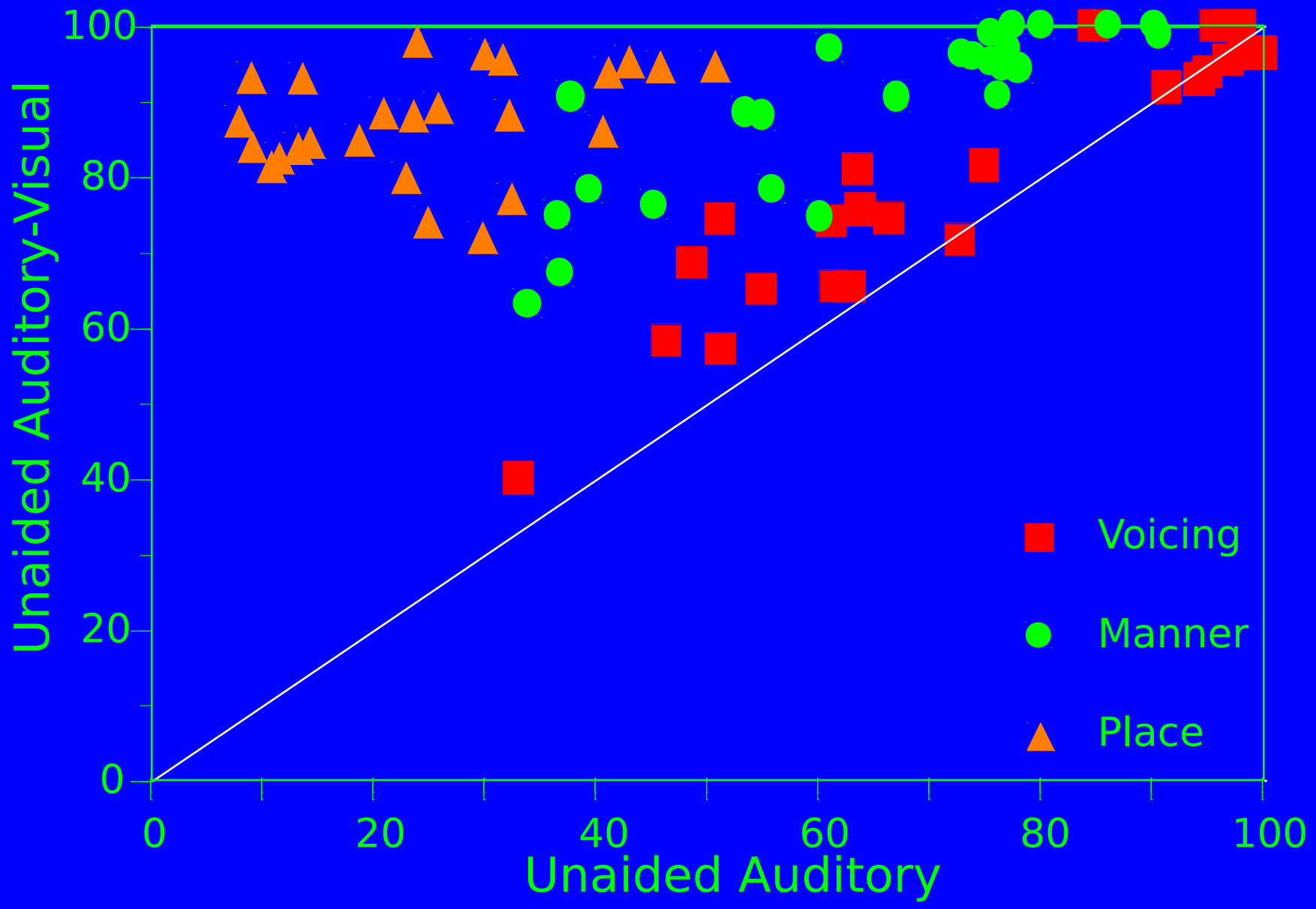


%Information Transmitted re: Total Information Receive

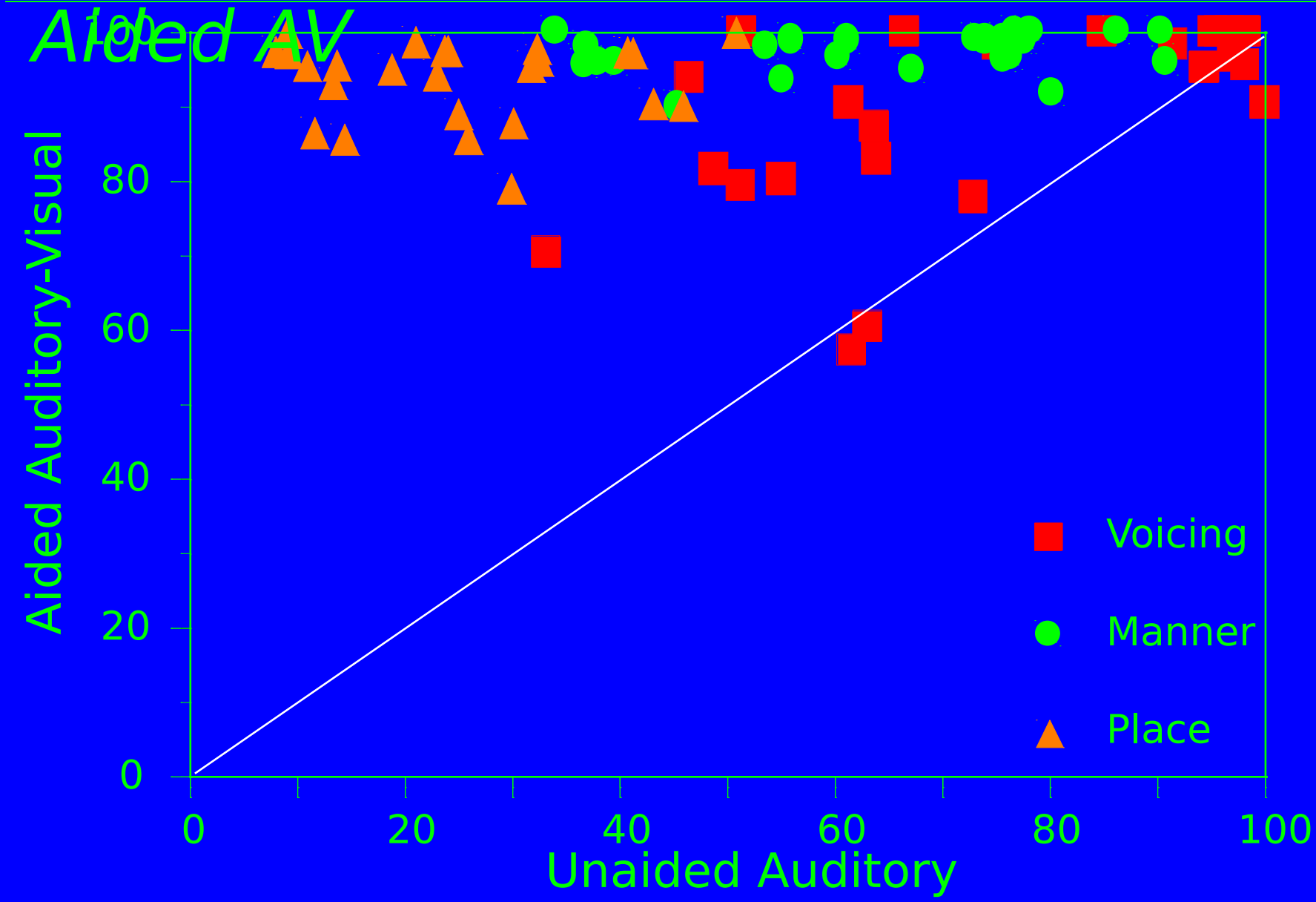
Speech Feature Distribution - Aided A



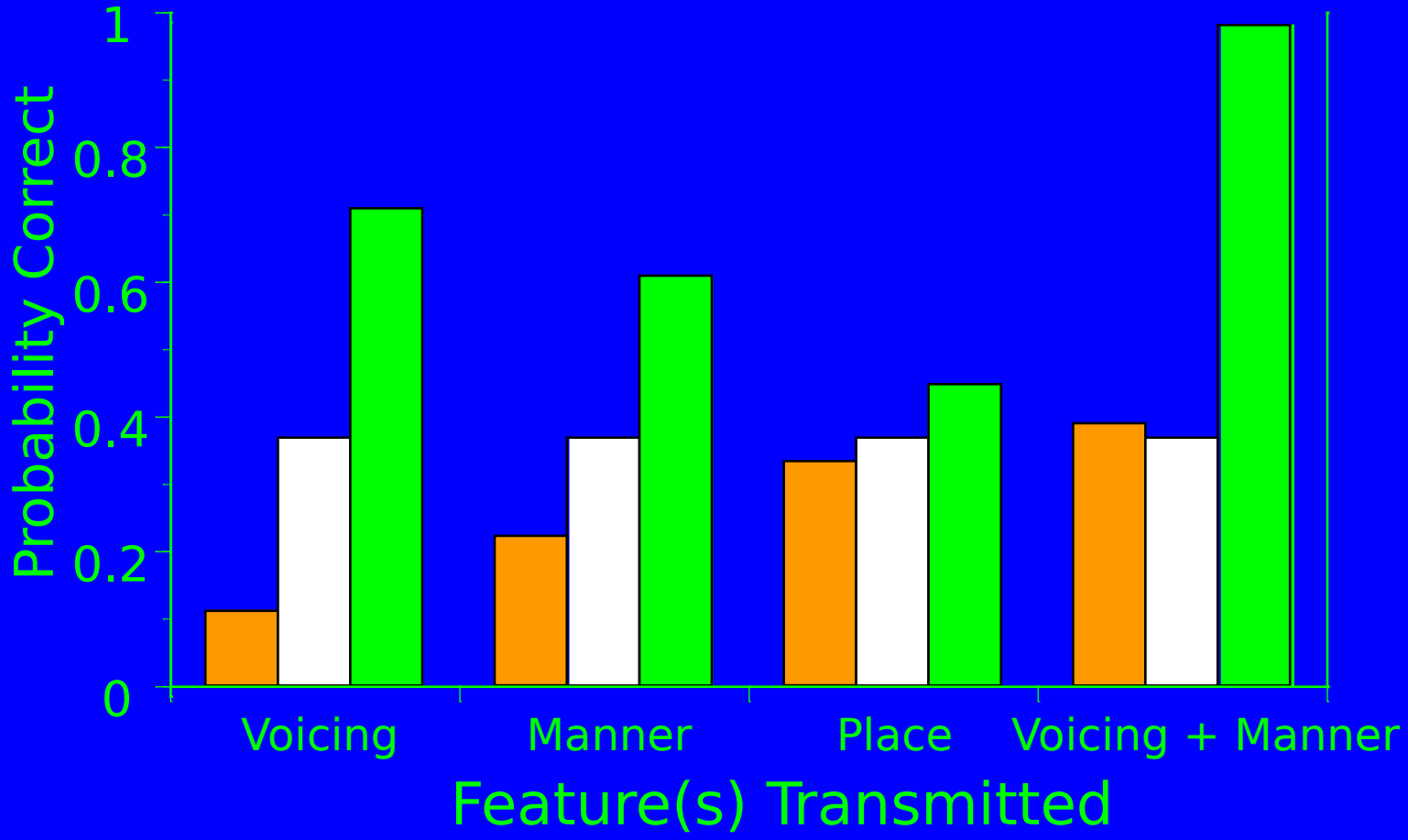
Speech Feature Distribution - Unaided



Speech Feature Distribution - Aided AV



Hypothetical Consonant Recognition - Perfect Feature Transmis



A

V

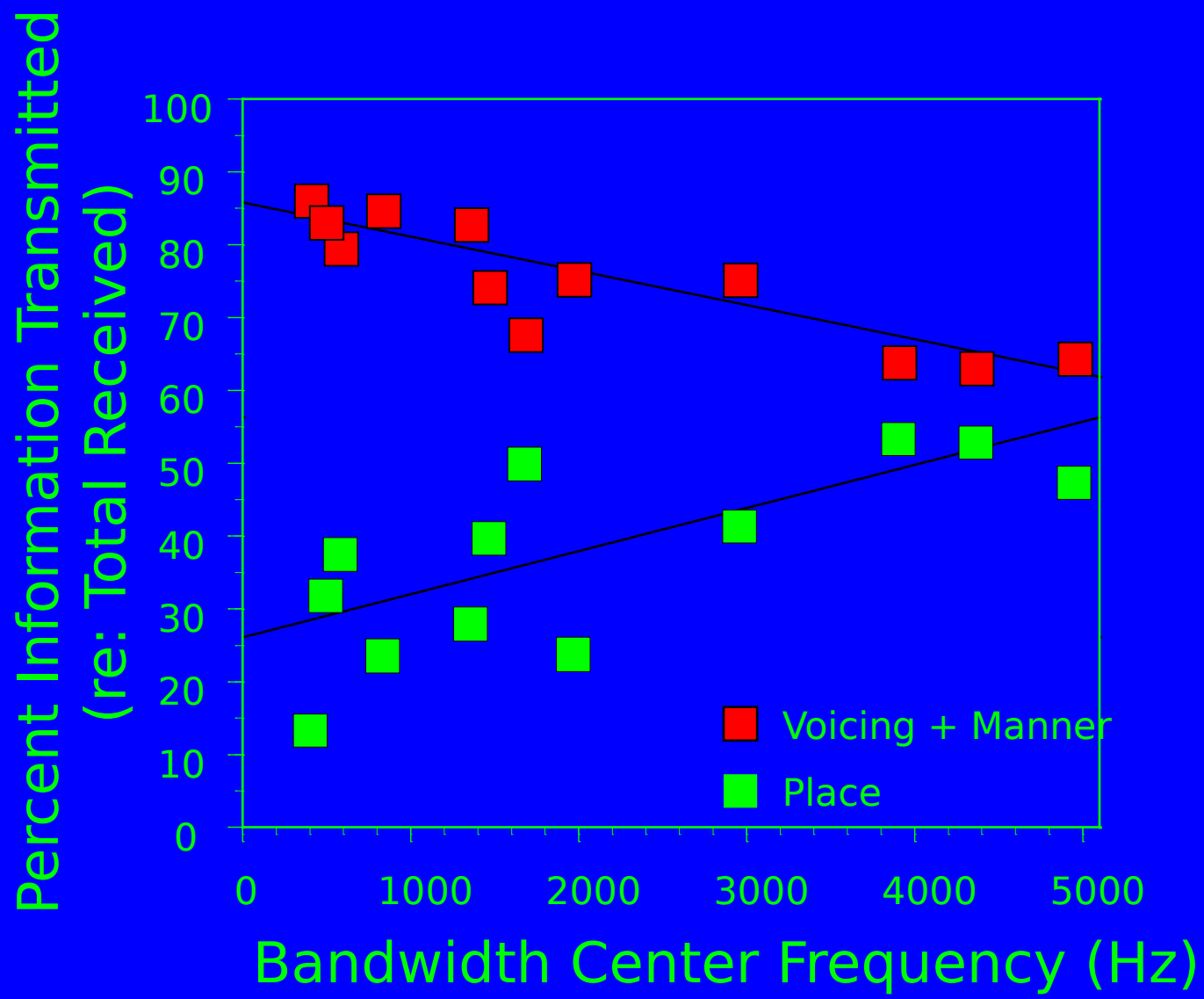
PRE

Auditory consonant recognition based on perfect transmission of indicated feature. Responses within each feature category were uniformly distributed.

Visual Consonant recognition. Data from 11 normally-hearing subjects (Grant and Walden, 1996)

Predicted AV consonant recognition based on PRE model of integration (Braida

Feature Distribution re: Center Freque



Summary: Hearing Aids and AV Speech

- Hearing aids enhance all speech features to some extent
  - Place cues are enhanced the most
  - Voicing cues are enhanced the least

- Visual cues enhance mostly place cues, with little support to manner and virtually no support to voicing

- Model predictions show that auditory place cues are redundant with speechreading and offer little benefit under AV conditions.

- Model predictions also show that AV speech recognition is best supported by auditory cues for voicing and manner

The Independence of Sensory Systems???

- ***Information is extracted independently from A and V modalities***
  - ***Early versus Late Integration***
  - ***Most models are "late integration" models***

The Independence of Sensory Systems ??

- Information is extracted independently from A and V modalities
  - Early versus Late Integration
  - Most models are "late integration" models

BUT

- ***Speechreading activates primary auditory cortex (cf. Sams et al., 1991)***

The Independence of Sensory Systems???

- Information is extracted independently from A and V modalities
  - Early versus Late Integration
  - Most models are "late integration" models

BUT

- Speechreading activates primary auditory cortex (cf. Sams et al., 1991)

- ***Population of neurons in cat Superior Colliculus respond only to bimodal input (cf. Stein and Meredith, 1993)***

Sensory Integration: Many Questions

- ***How is the auditory system modulated by visual speech activity?***
- ***What is the temporal window governing this interaction?***

The Effect of Speechreading on Masked Detection Thresholds for Spoken Sentences

Ken W. Grant and Philip Franz Seitz

Army Audiology and Speech Center
Walter Reed Army Medical Center
Washington, D.C. 20307, USA

http://www.wramc.amedd.army.mil/departments/aasc/avlab/
grant@tidalwave.net

Acknowledgments and Thanks

Technical Assistance

Philip Franz Seitz

Research Funding

National Institutes for Health

Coherence Masking Protection

- *CMP (Gordon, 1997a,b)*
  - Identification or classification of acoustic signals in noise is improved by adding additional signal components which are non-informative by themselves, provided that the relevant and non-relevant signal components “cohere” in some way (e.g., common fundamental frequency, common onset and offset, etc.)

Bimodal Coherence Masking Protection

- *BCMP (Grant and Seitz, 2000, Grant, 2001)*
  - Detection of speech in noise is improved by watching a talker (i.e., speechreading) as they produce the target speech signal, provided that the "*visible*" movement of the lips and acoustic amplitude envelope are highly correlated.

Bimodal Coherence

- *Acoustic/Optic Relationship*
- *Reduction of Signal Uncertainty*
  - in time domain
  - in spectral domain
- *Greater Neural Drive to Cognitive Centers*
  - neural priming - auditory cortex
  - bimodal cells in superior colliculus

Basic Paradigm for BCMP

- **Auditory-only speech detection**

speech

noise **noise**

- **Auditory-visual speech detection**

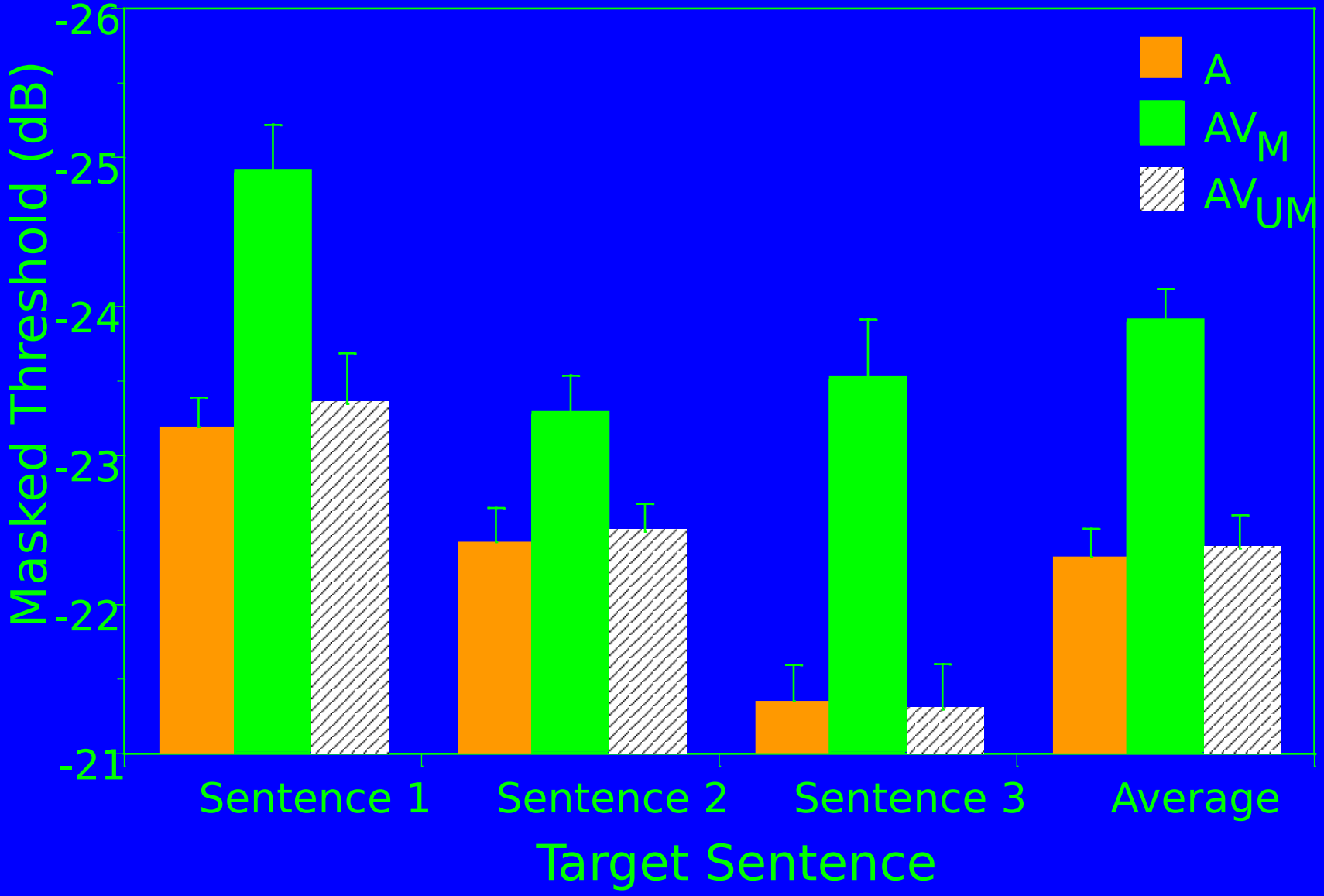
speech

noise **noise**

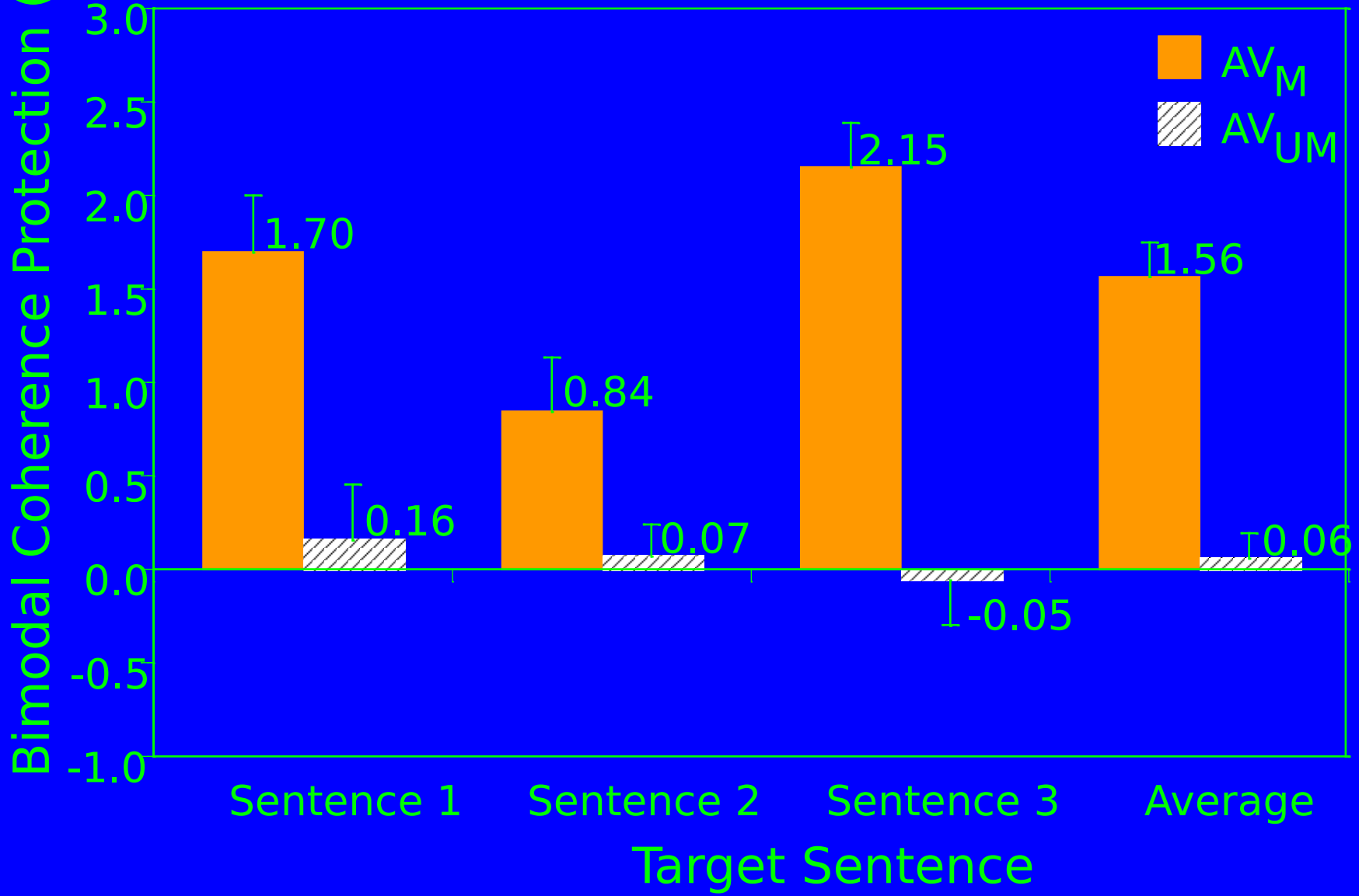
Experiment 1: Natural Speech

- 9 Interleaved Tracks - 2IFC Sentence Detection in Noise
  - 3 Audio target sentences - IEEE/Harvard Corpus
  - Same 3 target sentences with matched video
  - Same 3 target sentences with unmatched video (from 3 additional sentences)
  - White noise masker centered on target with random 100-300 ms extra onset and offset duration
- 10 Normal-Hearing Subjects

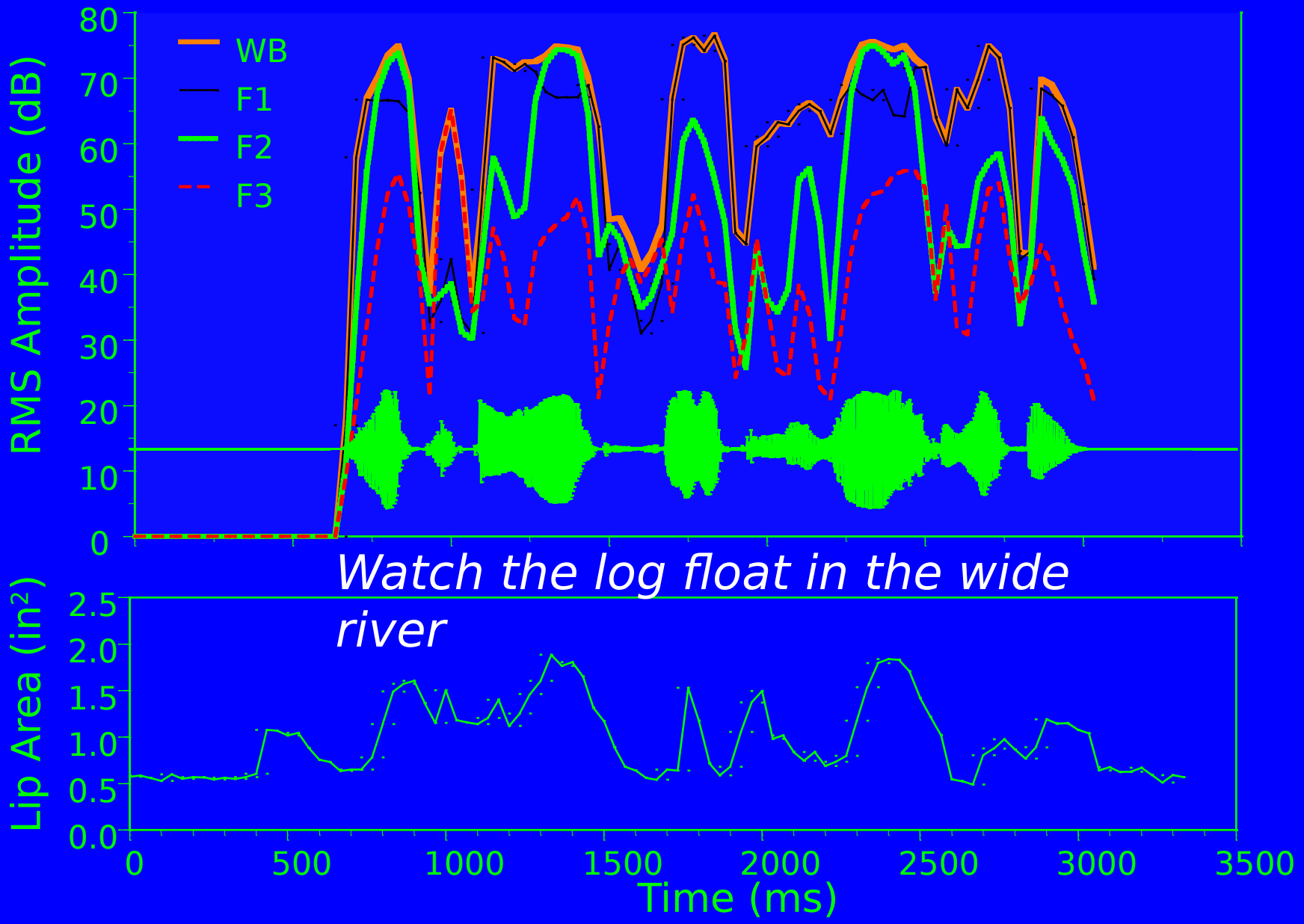
Congruent versus Incongruent Speech



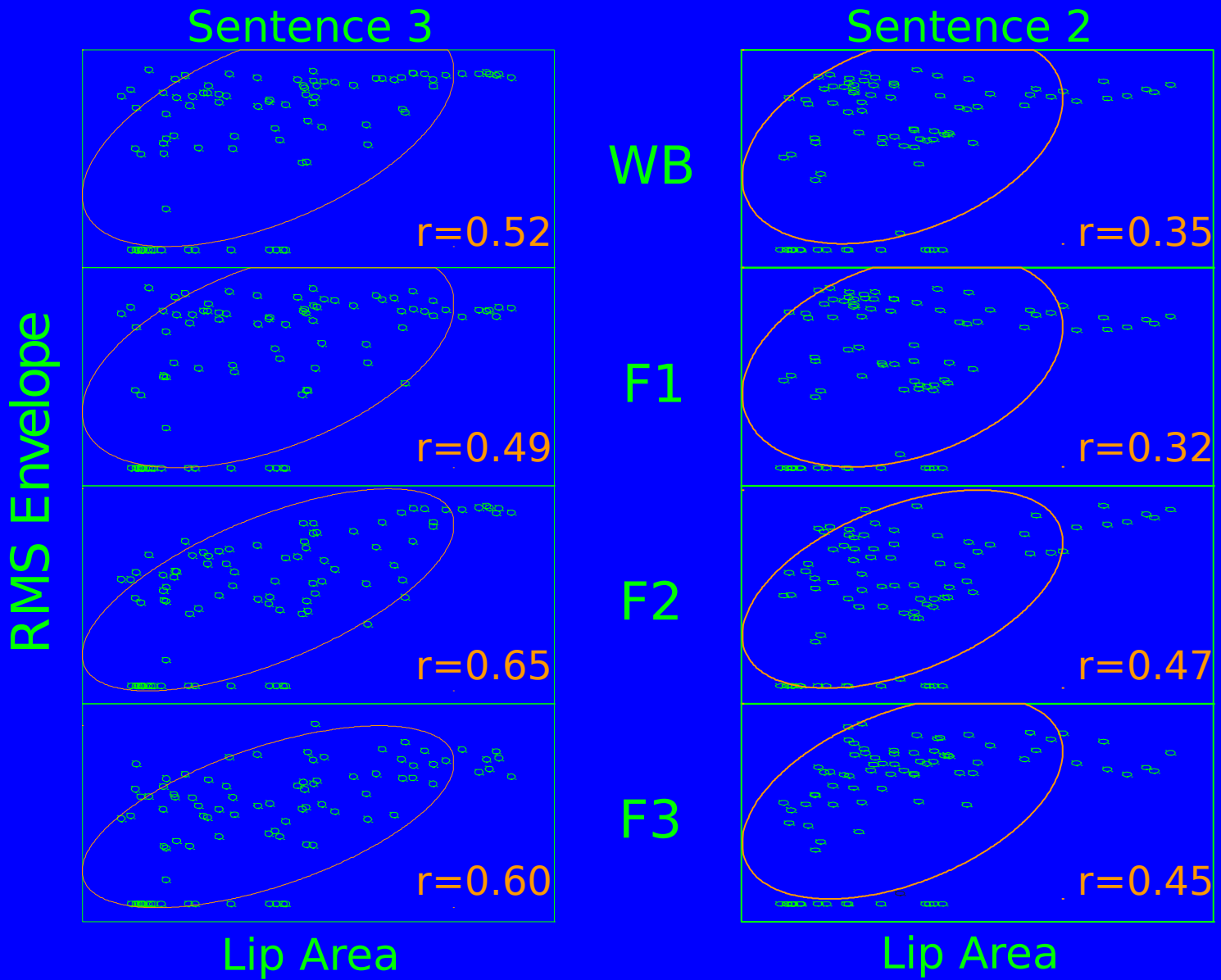
BCMP for Congruent and Incongruent Spee



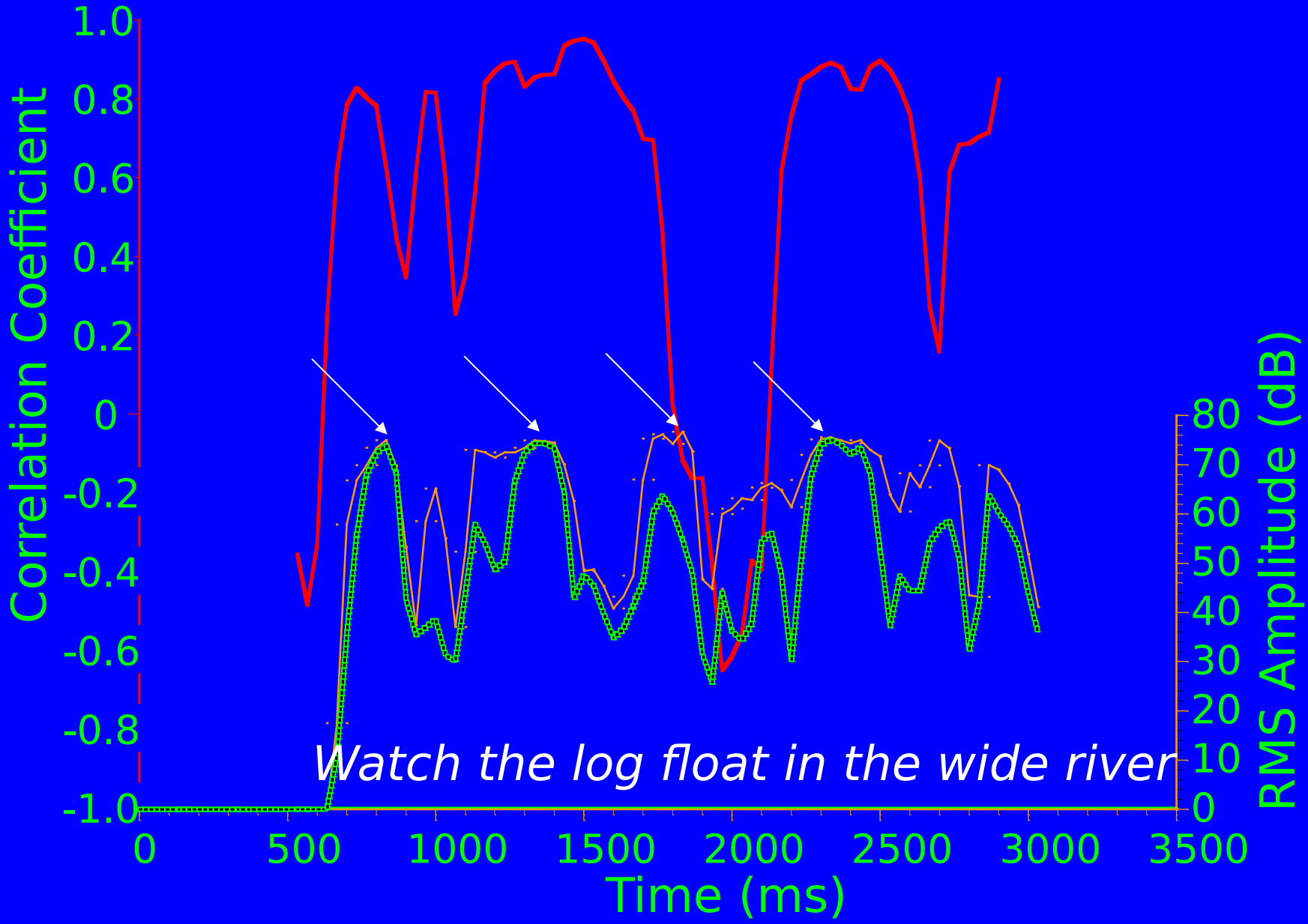
Acoustic Envelope and Lip Area Functions



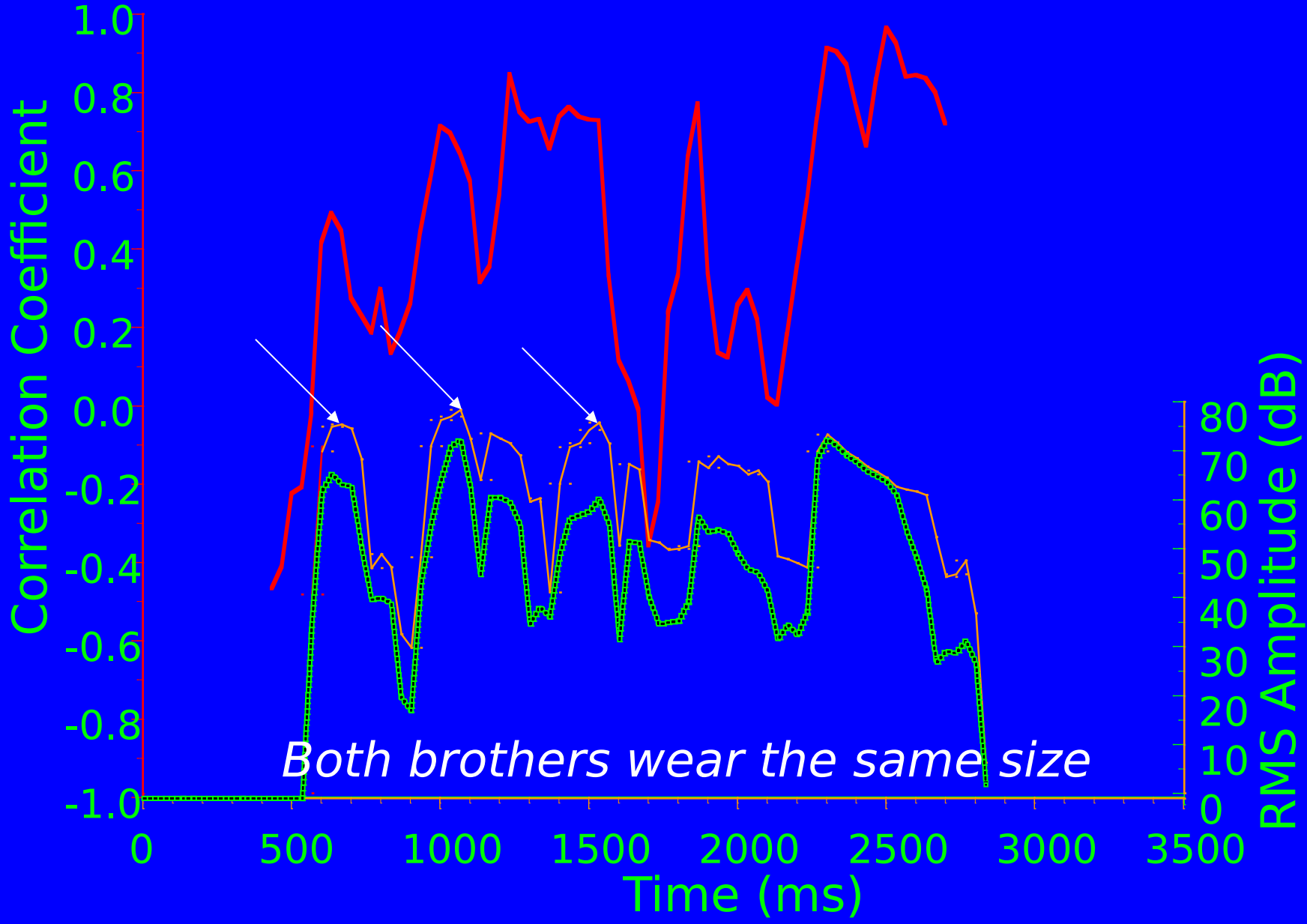
Cross Modality Correlation - Lip Area versus Amplitude Envelo



Local Correlations (Lip Area versus F2-Amplitude Envelop

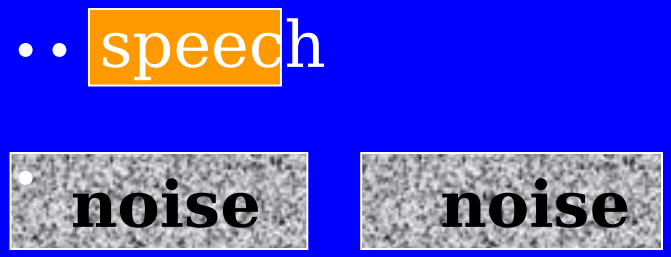


Local Correlations (Lip Area versus F2-Amplitude Envelop

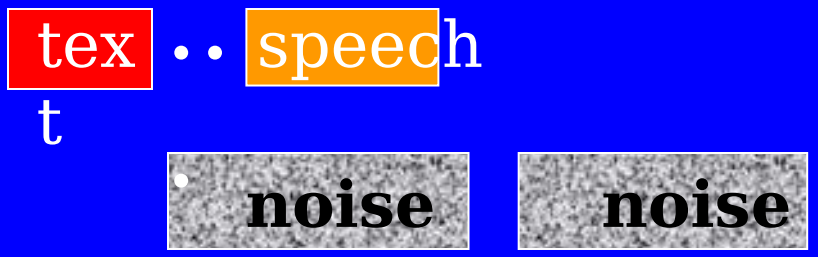


Methodology for Orthographic BCMP

- **Auditory-only speech detection**



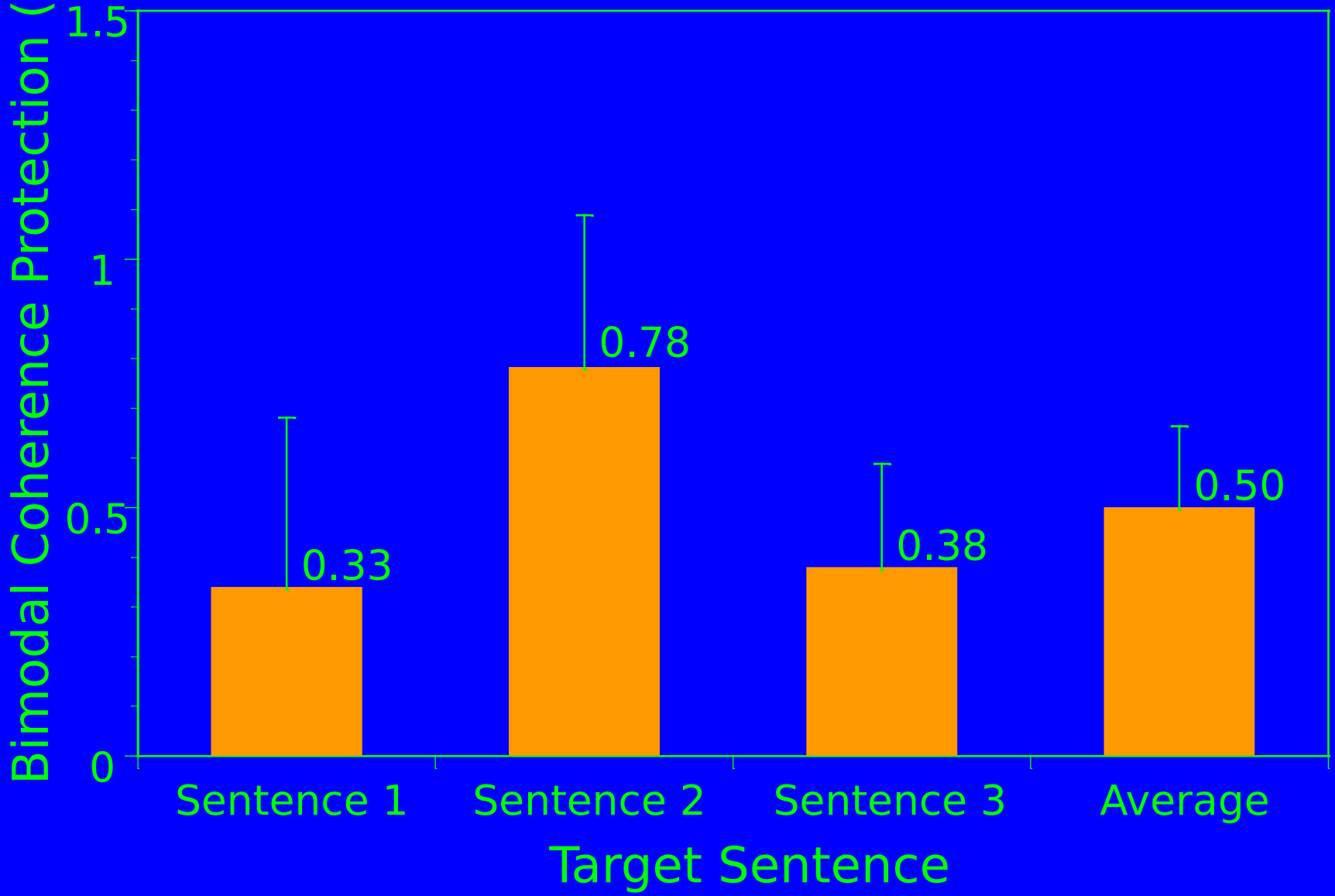
- **Auditory + orthographic speech detection**



Experiment 2: Orthography

- 6 Interleaved Tracks - 2IFC Sentence Detection in Noise
  - 3 Audio target sentences - IEEE/Harvard Corpus
  - Same 3 target sentences with matched orthography
  - White noise masker centered on target with random 100-300 ms extra onset and offset duration
- 6 Normal-Hearing Subjects (from EXP 1)

BCMP for Orthographic Speech



Methodology for Filtered BCMP

F1 (100-800 Hz) F2 (800-2200 Hz)

- Auditory-only detection of filtered speech

  - speech

  - **noise** **noise**

- Auditory-visual detection of filtered speech

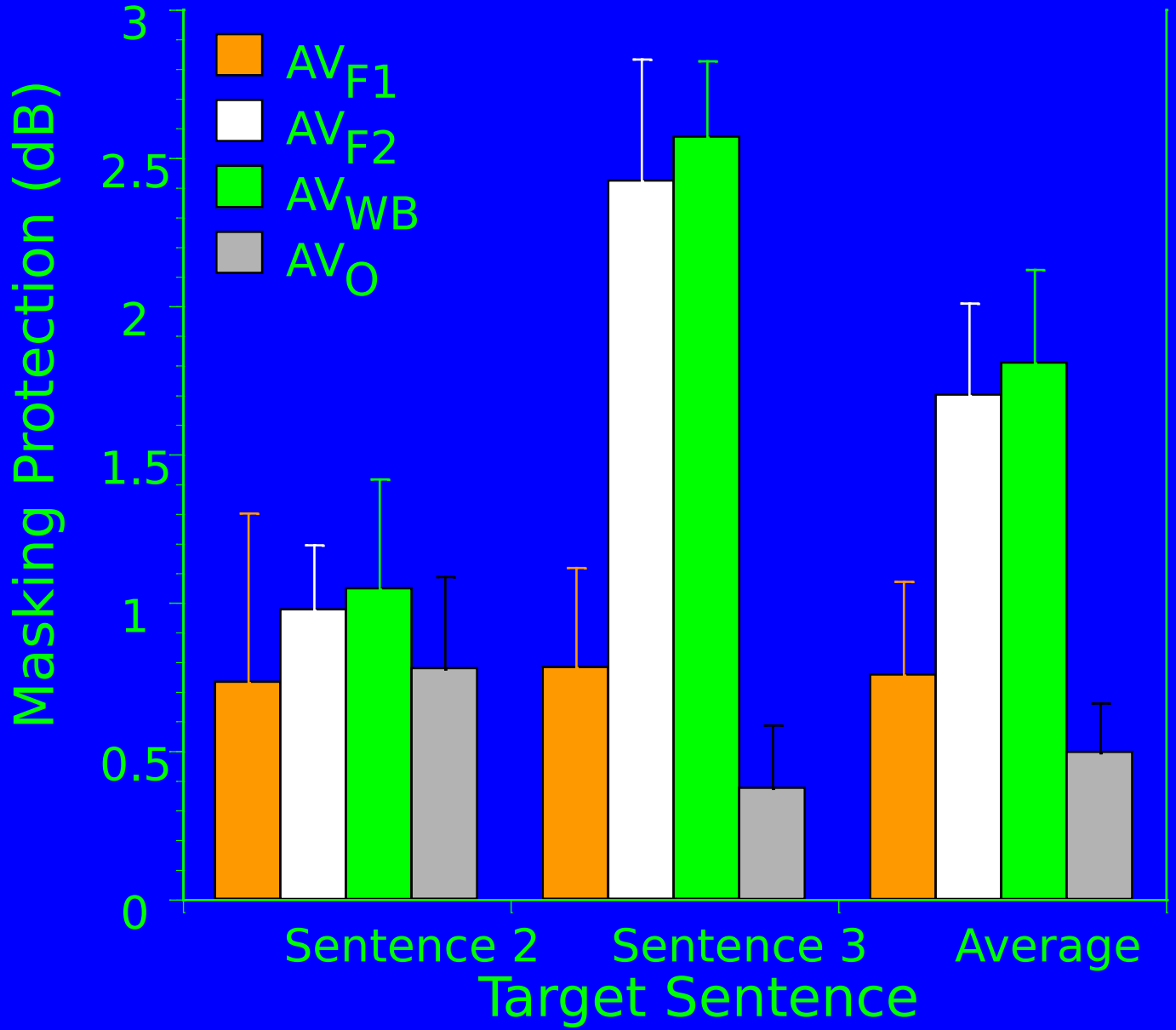
  - speech

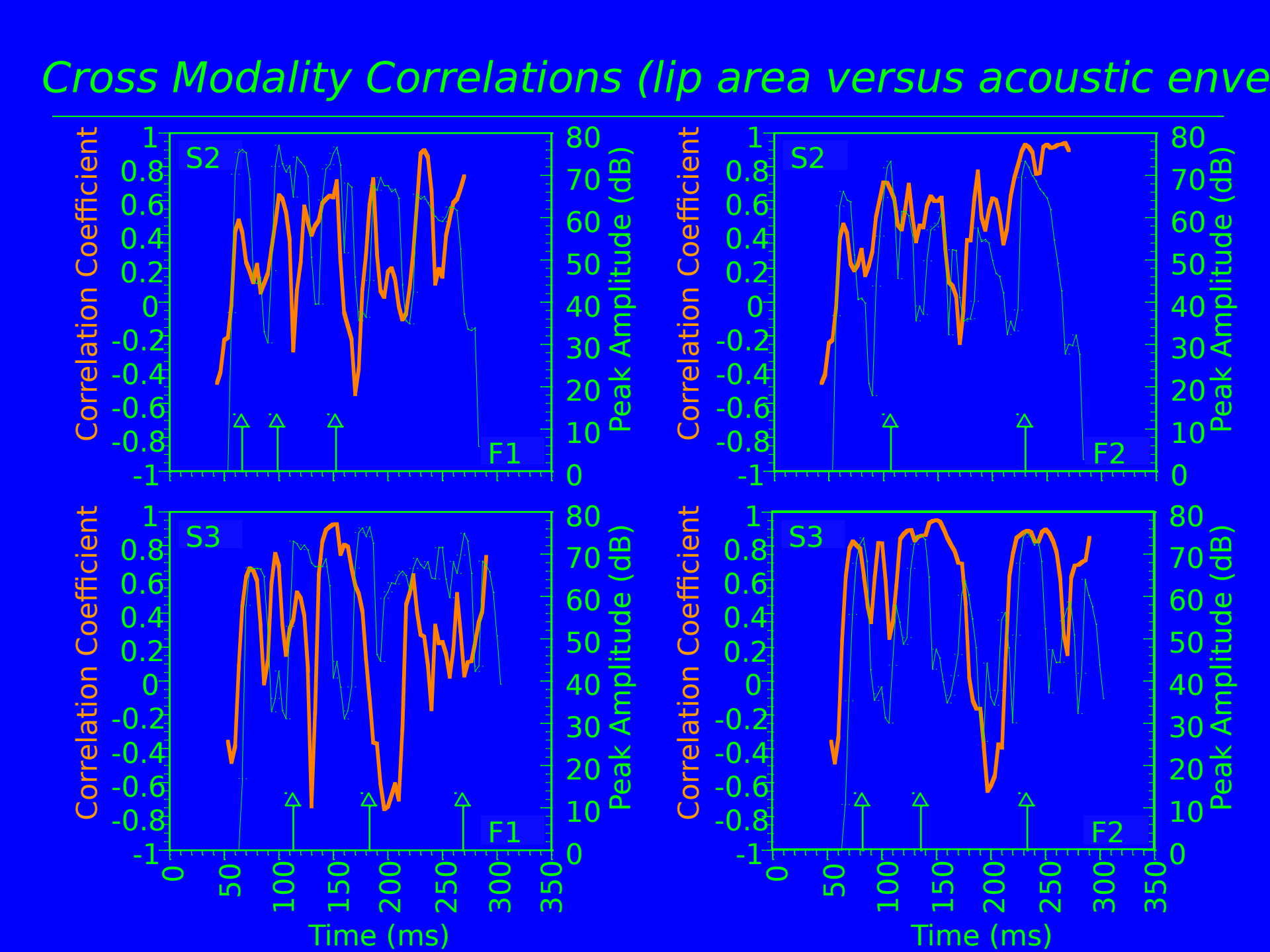
  - **noise** **noise**

Experiment 3: Filtered Speech

- 6 Interleaved Tracks - 2IFC Sentence Detection in Noise
  - 2 Audio target sentences - IEEE/Harvard Corpus - each filtered into F1 and F2 bands
  - Same 2 target sentences that produced largest BCMP
  - White noise masker centered on target with random 100-300 ms extra onset and offset duration
- 6 Normal-Hearing Subjects (from EXP 2)

BCMP for Filtered Speech




Cross Modality Correlations (lip area versus acoustic enve
S2
F1
S2
F2
S3
F1
S3
F2
Correlation Coefficient
Peak Amplitude (dB)
Time (ms)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
80
70
60
50
40
30
20
10
0
0
50
100
150
200
250
300
350

BCMP – Summary

- ***Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- ***Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics
- ***Providing listeners with explicit (orthographic) knowledge of the identity of the target sentence reduces speech detection thresholds by about 0.5 dB, independent of the specific target sentence***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics
- Providing listeners with explicit (orthographic) knowledge of the identity of the target sentence reduces speech detection thresholds by about 0.5 dB, independent of the specific target sentence
- ***Manipulating the degree of coherence between area of mouth opening and acoustic envelope by filtering the target speech has a direct effect on BCMP***

Temporal Window for Auditory-Visual Integra

Speech Intelligibility Derived from Asynchronous Processing of Auditory-Visual Information

Ken W. Grant
Army Audiology and Speech Center
Walter Reed Army Medical Center
Washington, D.C. 20307, USA
http://www.wramc.amedd.army.mil/departments/aasc/avlab
grant@tidalwave.net

Steven Greenberg
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
http://www.icsi.berkeley.edu/~steveng
steveng@icsi.berkeley.edu

Acknowledgments and Thanks

Technical Assistance

Takayuki Arai, Rosaria Silipo

Research Funding

U.S. National Science Foundation

Spectral-Temporal Integration

- Time-Course of Spectral-Temporal Integration
  - ***Within Modality***
    - 4 narrow spectral slits
    - delay two middle bands with respect to outer bands
  - ***Across Modality***
    - 2 narrow spectral slits (2 outer bands)
    - delay audio with respect to video

Auditory-Visual Asynchrony - Paradigm

Video of spoken (Harvard/IEEE) sentences, presented in tandem with asparse spectral representation (low- and high-frequency slits) of the same material

Baseline Condition
SYNCHRONOUS A/V

Auditory-Visual Integration - Preview

When the AUDIO signal LEADS the VIDEO, there is a progressive decline in intelligibility, similar to that observed for audio-alone signals

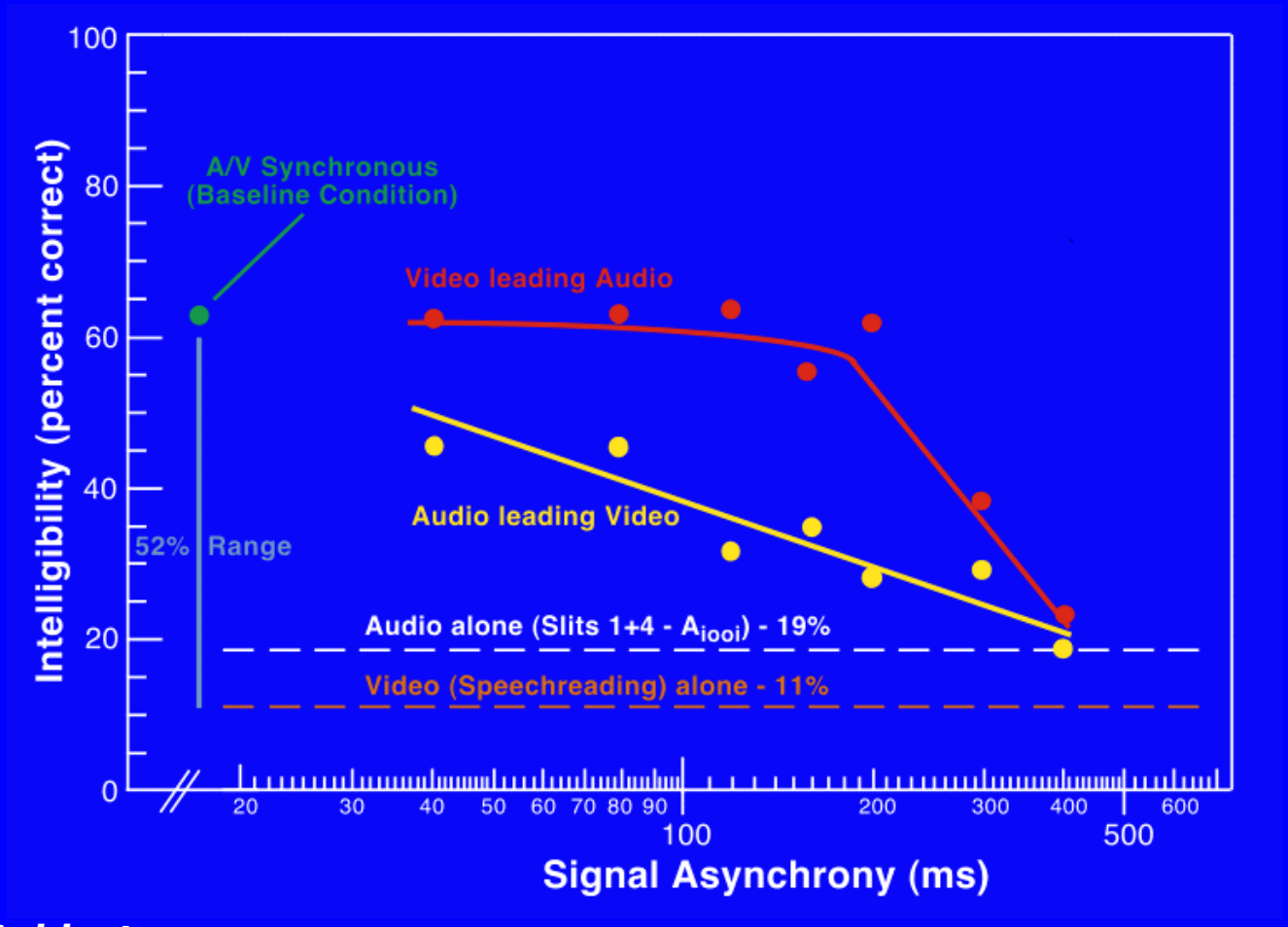
When the VIDEO signal LEADS the AUDIO, intelligibility is preserved for asynchrony intervals as long as 200 ms

Why? Why? Why?

We'll return to these data shortly

But first, let's take a look at audio-alone speech intelligibility data in order to gain some perspective on the audio-visual case

The audio-alone data come from earlier studies by



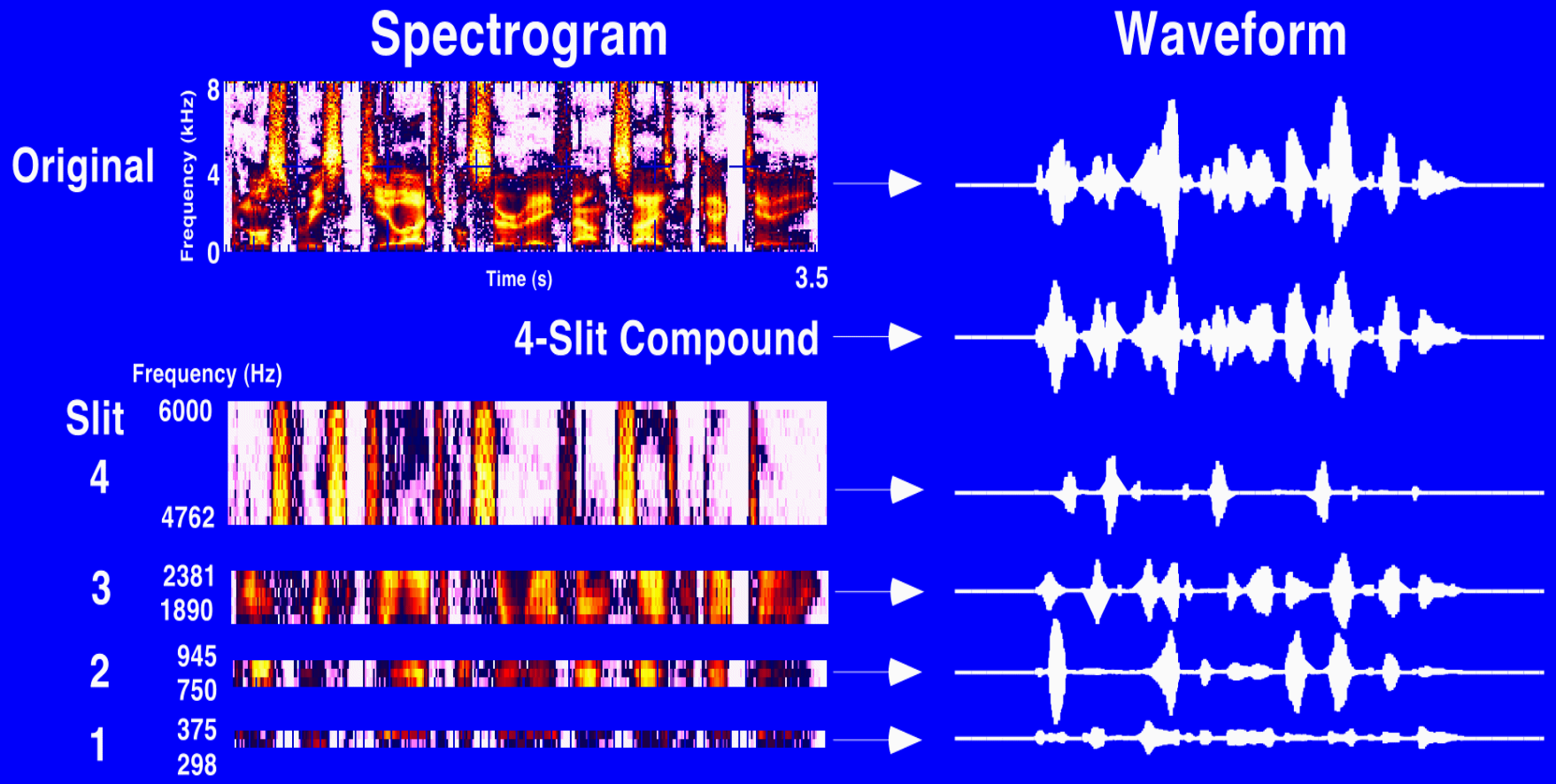
AUDIO-ALONE EXPERIMENTS

Audio (Alone) Spectral Slit Paradigm

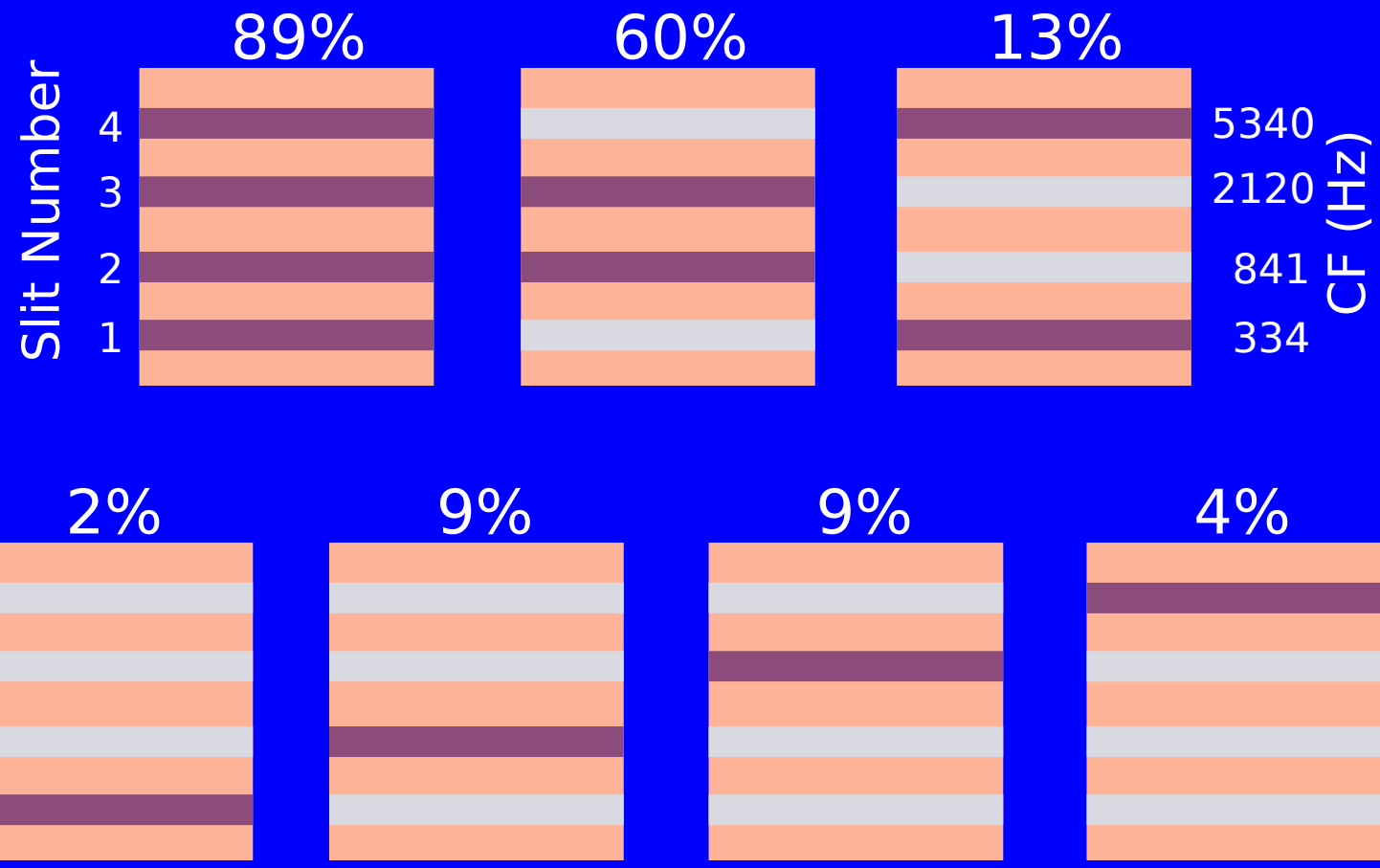
The edge of each slit was separated from its nearest neighbor by an octave

Can listeners decode spoken sentences using just four narrow (1/3 octave) channels ("slits") distributed across the spectrum? – YES (cf. next slide)

What is the intelligibility of each slit alone and in combination with others?



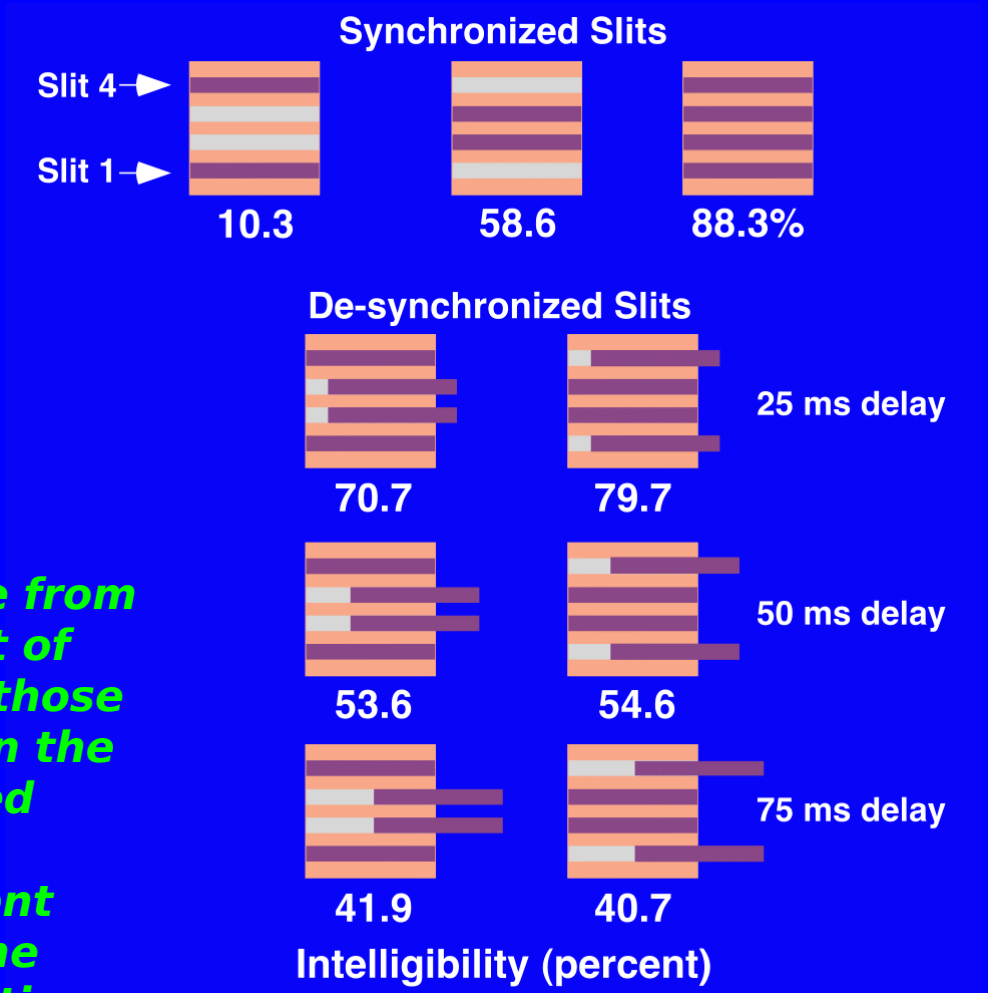
Word Intelligibility - Single and Multiple Sl



Slit Asynchrony Affects Intelligibility

Desynchronizing the slits by more than 25 ms results in a significant decline in intelligibility

The effect of asynchrony on intelligibility is relatively symmetrical

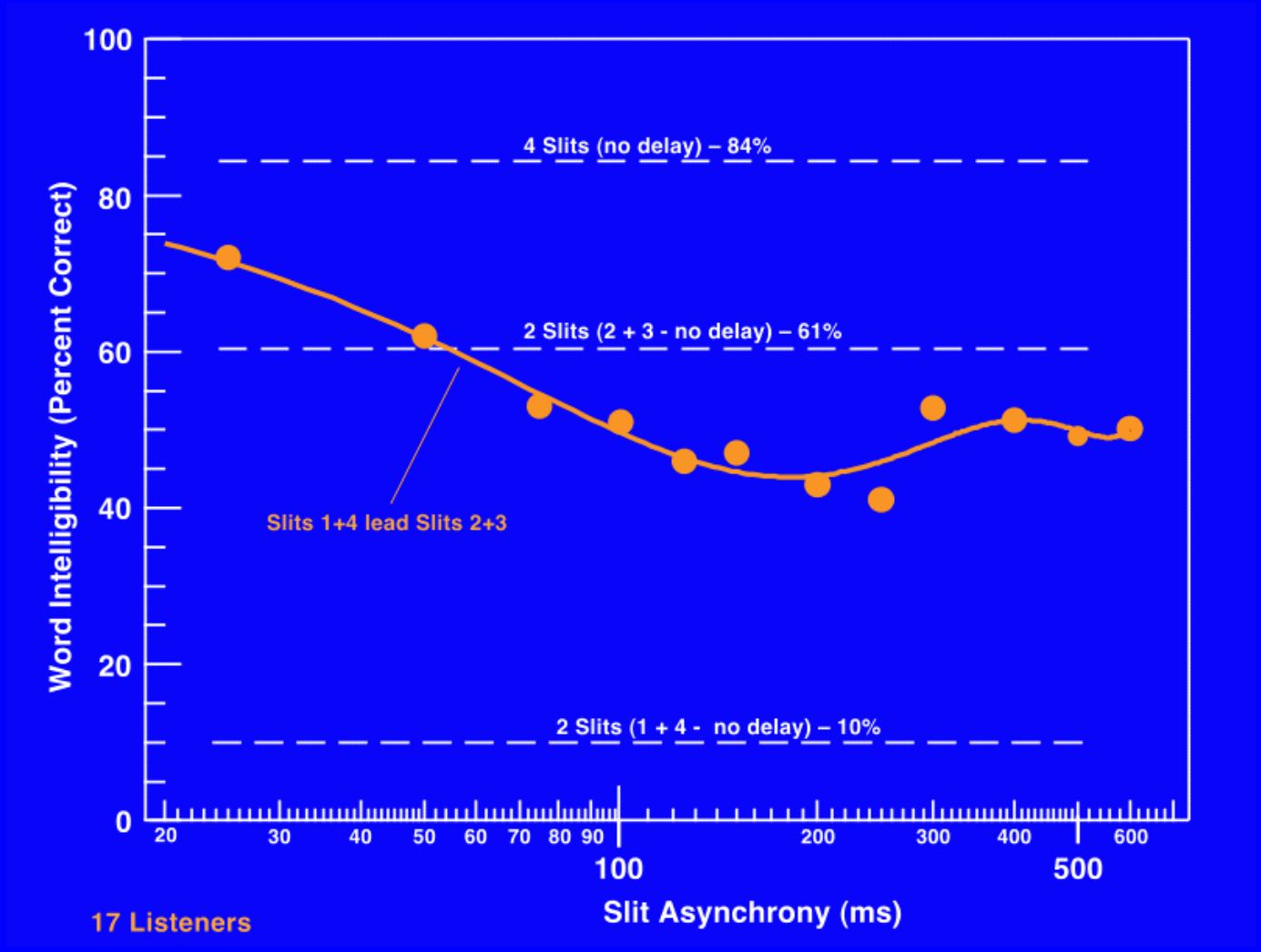


These data are from a different set of subjects than those participating in the study described earlier - hence slightly different numbers for the baseline conditions

Intelligibility and Slit Asynchrony

Desynchronizing the two central slits relative to the lateral ones has a pronounced effect on intelligibility

Asynchrony greater than 50 ms results in intelligibility lower than baseline

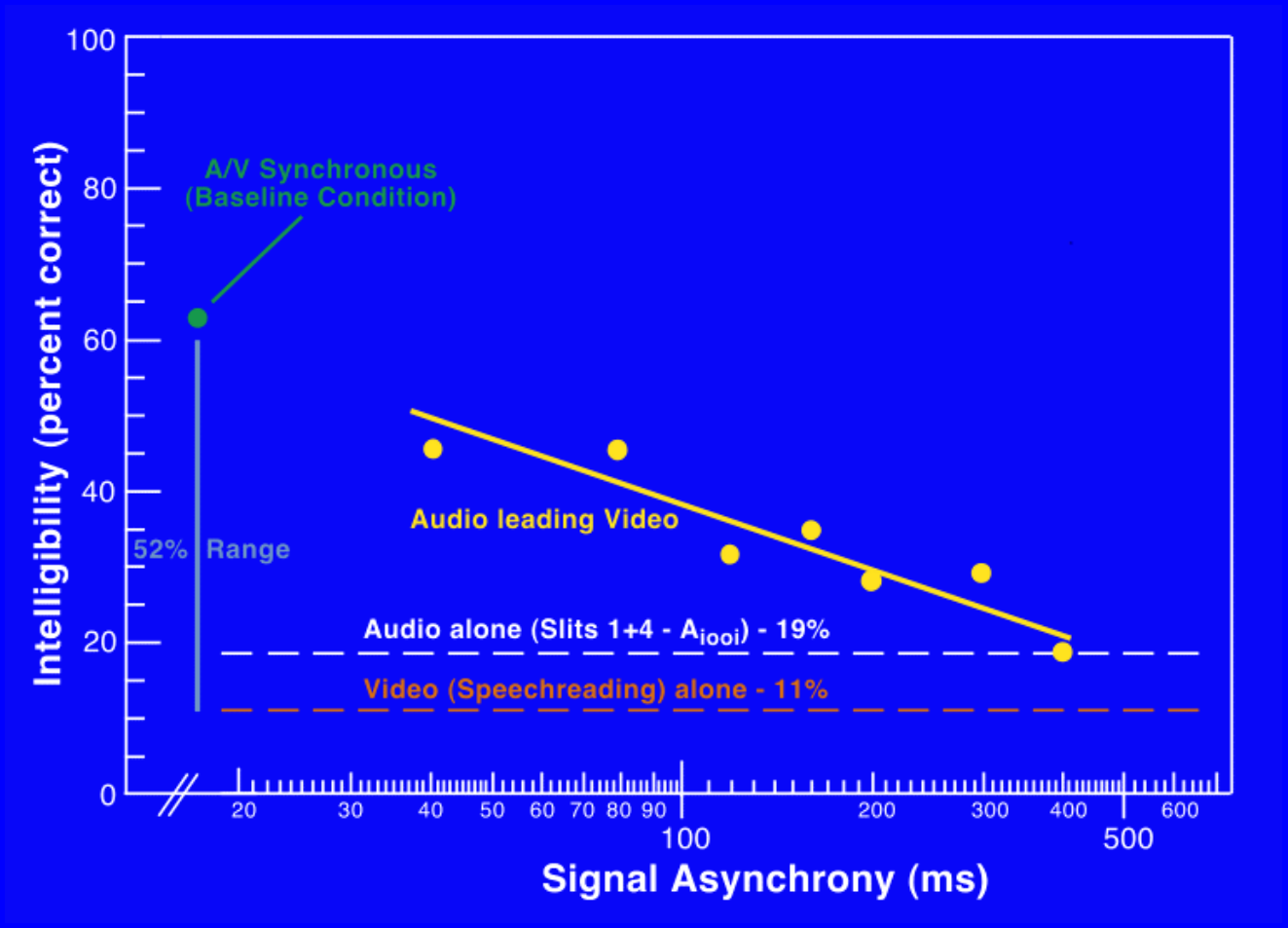


AUDIO-VISUAL EXPERIMENTS

Focus on Audio-Leading-Video Conditions

When the AUDIO signal LEADS the VIDEO, there is a progressive decline in intelligibility, similar to that observed for audio-alone signals

These data are next compared with data from the previous slide to illustrated the similarity in the slope of the function

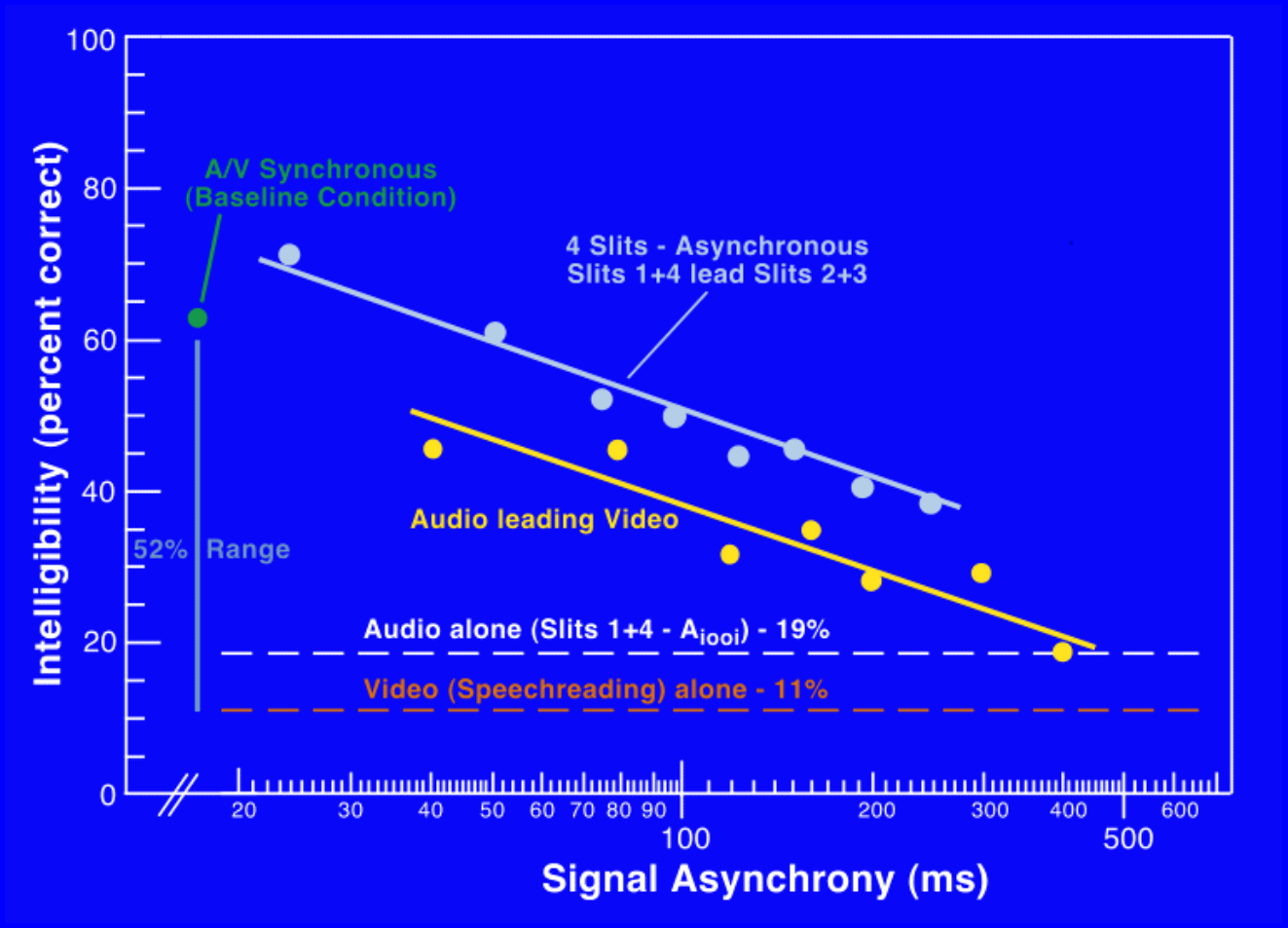


Comparison of A/V and Audio-Alone Data

The decline in intelligibility for the audio-alone condition is similar to that of the audio-leading-video condition

Such similarity in the slopes associated with intelligibility for both experiments suggest that the underlying mechanisms may be similar

The intelligibility of the audio-alone signals is higher than the A/V signals due to slits 2+3 being highly intelligible by themselves

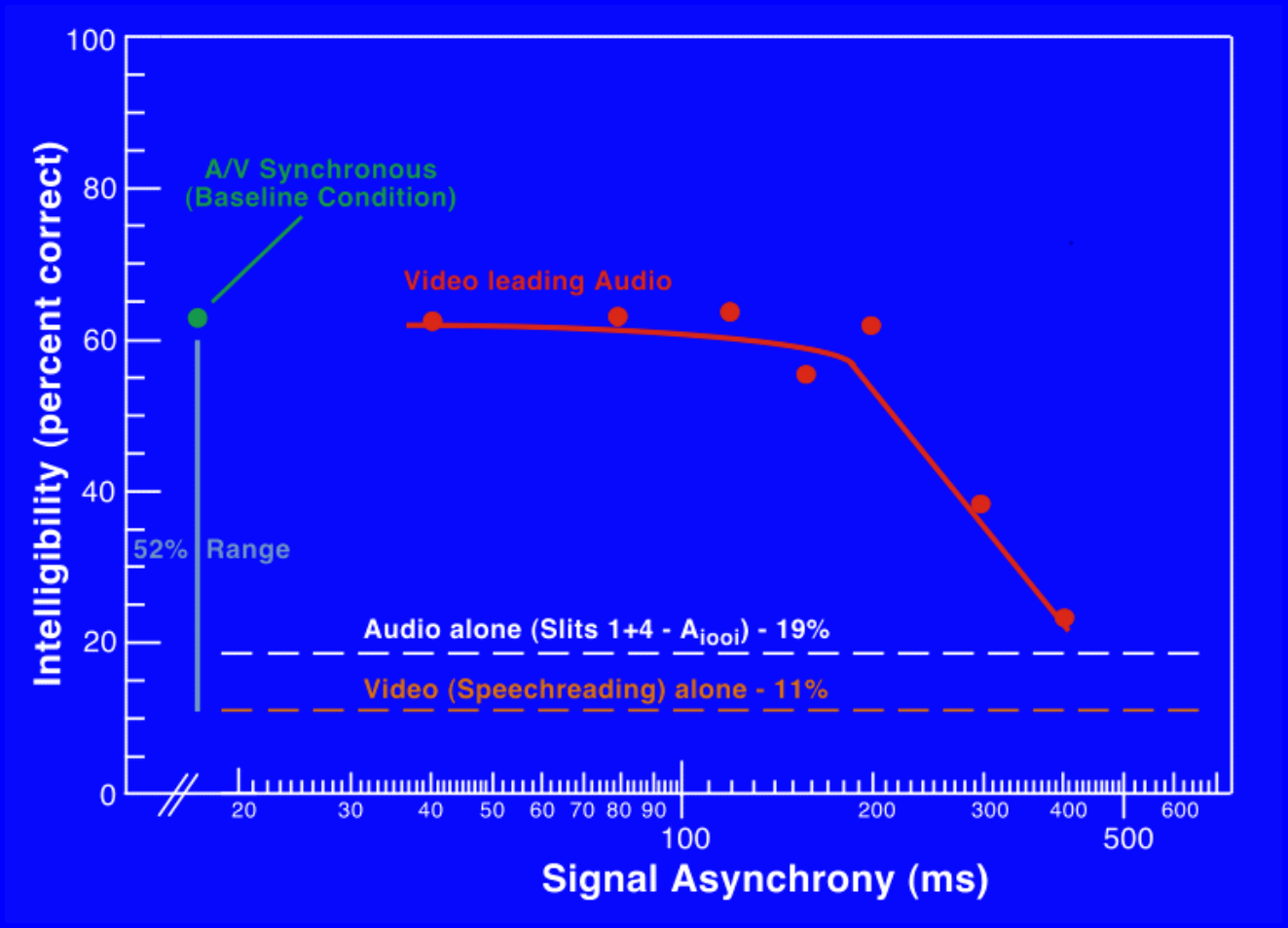


Focus on Video-Leading-Audio Conditions

When the VIDEO signal LEADS the AUDIO, intelligibility is preserved for asynchrony intervals as large as 200 ms

These data are rather strange, implying some form of "immunity" against intelligibility degradation when the video channel leads the audio

We'll consider a variety of interpretations in a few minutes

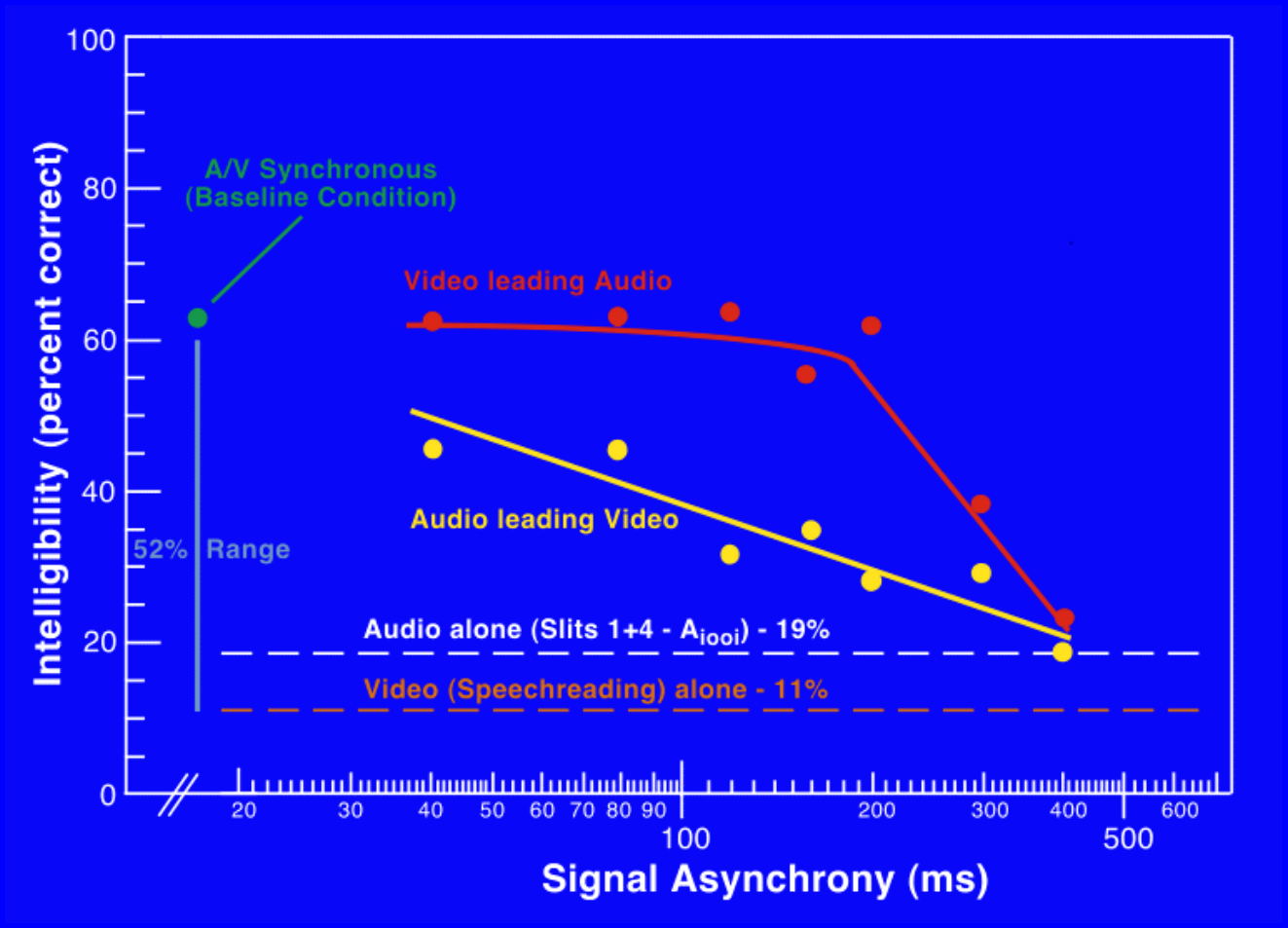


Auditory-Visual Integration

The slope of intelligibility-decline associated with the video-leading-audio conditions is rather different from the audio-leading-video conditions

WHY? WHY? WHY?

There are several interpretations of these data - we'll consider several on the following slide



INTERPRETATION OF THE DATA

Possible Interpretations of the Data – 1

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

Possible Interpretations of the Data – 1

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

***Visual information is processed in the brain much more slowly than auditory information* (difference in first spike latency in V1 versus A1 is about 30-50 ms)**

Possible Interpretations of the Data – 1

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

Visual information is processed in the brain much more slowly than auditory information (difference in first spike latency in V1 versus A1 is about 30-50 ms)

Some problems with this interpretation ….

Possible Interpretations of the Data – 1

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

Visual information is processed in the brain much more slowly than auditory information (difference in first spike latency in V1 versus A1 is about 30-50 ms)

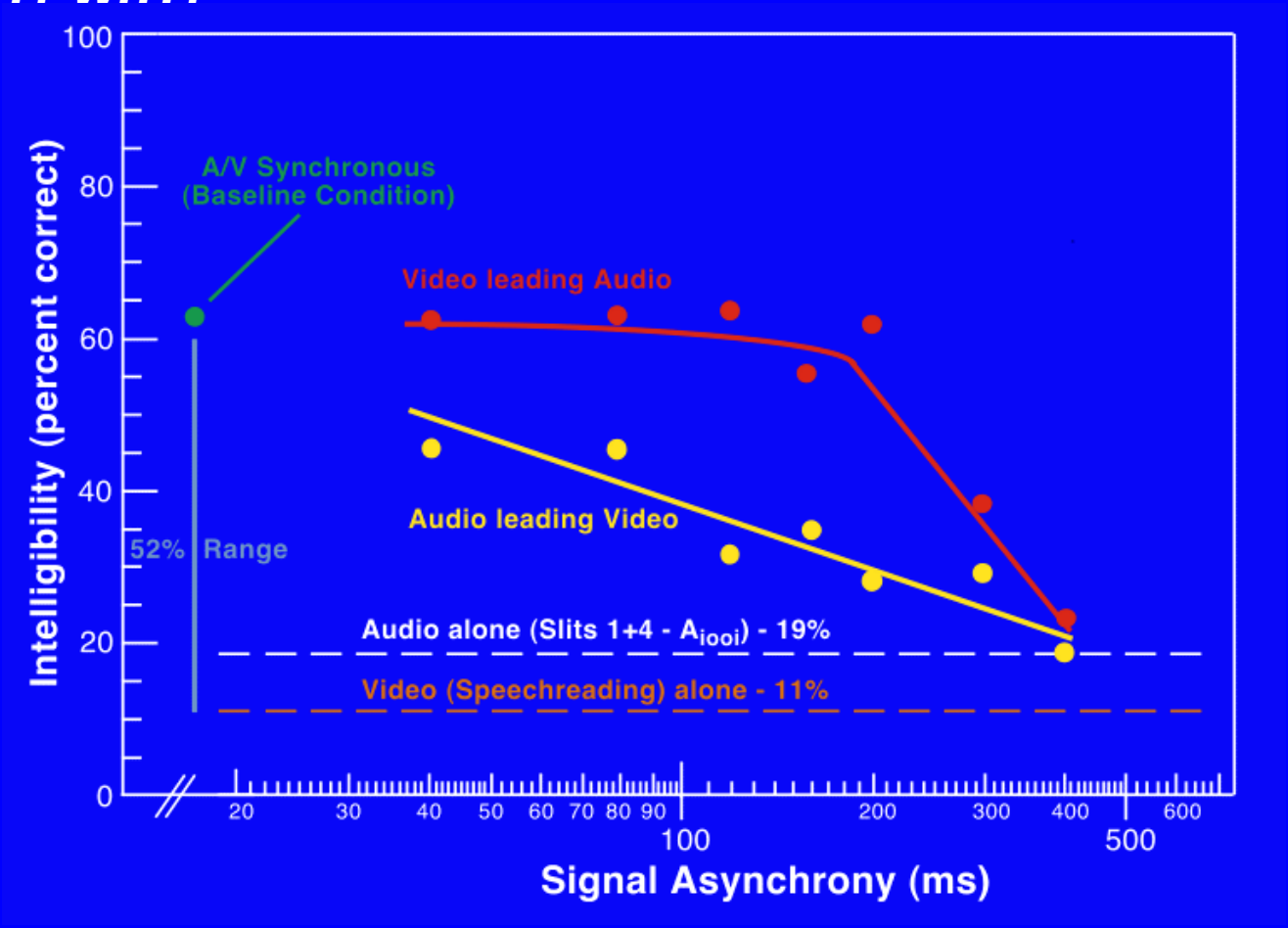
Some problems with this interpretation ….

Simple differences in neural conduction times may account for a preference for visual leading information BUT can not account for asymmetry in temporal window

Auditory-Visual Integration

The slope of intelligibility-decline associated with the video-leading-audio conditions is rather different from the audio-leading-video conditions

WHY? WHY? WHY?



Possible Interpretations of the Data – 2

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

Possible Interpretations of the Data – 2

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ....

There is certain information in the video component of the speech signal (and in the mid-frequency audio channels) that evolves over a relatively long interval of time (e.g., 200 ms) and is thus relatively immune to asynchronous combination with information contained in the audio channel. What might this information be?

Possible Interpretations of the Data – 2

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

There is certain information in the video component of the speech signal (and in the mid-frequency audio channels) that evolves over a relatively long interval of time (e.g., 200 ms) and is thus relatively immune to asynchronous combination with information contained in the audio channel. What might this information be?

The visual component of the speech signal is most closely associated with place-of-articulation information (cf. Grant and Walden, 1996)

Possible Interpretations of the Data – 2

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

There is certain information in the video component of the speech signal (and in the mid-frequency audio channels) that evolves over a relatively long interval of time (e.g., 200 ms) and is thus relatively immune to asynchronous combination with information contained in the audio channel. What might this information be?

The visual component of the speech signal is most closely associated with place-of-articulation information (Grant and Walden, 1996)

In the (audio) speech signal, place-of-articulation information usually evolves over two or three phonetic segments (i.e., a syllable in length)

Possible Interpretations of the Data – 2

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

There is certain information in the video component of the speech signal that evolves over a relatively long interval of time (e.g., 200 ms) and is thus relatively immune to asynchronous combination with information contained in the audio channel. What might this information be?

The visual component of the speech signal is most closely associated with place-of-articulation information (Grant and Walden, 1996)

In the (audio) speech signal, place-of-articulation information usually evolves over two or three phonetic segments (i.e., a syllable in length)

BUT ….. the data imply that the modality arriving first determines the mode (and hence the time constant of processing) for combining information across sensory channels

Possible Interpretations of the Data – 3

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

Possible Interpretations of the Data – 3

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ....

The brain has evolved under conditions where the visual signal arrives prior to the audio signal, but where the time disparity between the two modalities varies from situation to situation

Possible Interpretations of the Data – 3

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

The brain has evolved under conditions where the visual signal arrives prior to the audio signal, but where the time disparity between the two modalities varies from situation to situation

Therefore, processing of A and V information is tolerant of asynchronous arrival times so long as the visual information leads the auditory information

Possible Interpretations of the Data – 3

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ….

The brain has evolved under conditions where the visual signal arrives prior to the audio signal, but where the time disparity between the two modalities varies from situation to situation

Therefore, processing of A and V information is tolerant of asynchronous arrival times so long as the visual information leads the auditory information

BUT… Why the 200 ms value? Given the speed of light and sound (cf. 1129 ft/sec), 200 ms would suggest that we can process AV stimuli at 225.8 ft. That's may be a bit much!

Possible Interpretations of the Data – 3

The video-leading-audio conditions are more robust in the face of asynchrony than audio-leading-video conditions because ....

The brain has evolved under conditions where the visual signal arrives prior to the audio signal, but where the time disparity between the two modalities varies from situation to situation

Therefore, processing of A and V information is tolerant of asynchronous arrival times so long as the visual information leads the auditory information

BUT... Why the 200 ms value? Given the speed of light and sound (cf. 1129 ft/sec), 200 ms would suggest that we can process AV stimuli at 225.8 ft. That's may be a bit much!

In addition to speed differences, we also must consider that speech movements precede audio output by variable amounts. The limit of 200 ms may reflect a boundary associated with syllabic length units, over which integration is probably not beneficial.

SUMMARY

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

When the audio signal leads the video, intelligibility falls off rapidly as a function of modality asynchrony

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

When the audio signal leads the video, intelligibility falls off rapidly as a function of modality asynchrony

When the video signal leads the audio, intelligibility is maintained for asynchronies as long as 200 ms

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

When the audio signal leads the video, intelligibility falls off rapidly as a function of modality asynchrony

When the video signal leads the audio, intelligibility is maintained for asynchronies as long as 200 ms

There are many potential interpretations of the data

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

When the audio signal leads the video, intelligibility falls off rapidly as a function of modality asynchrony

When the video signal leads the audio, intelligibility is maintained for asynchronies as long as 200 ms

There are many potential interpretations of the data

The simplest explanation is also the most ecologically valid: we prefer conditions where visual information leads audio information because that's what we're confronted with in the real world. But there are still questions remaining.

Audio-Video Integration – Summary

Spectrally sparse audio and speech-reading information provide minimal intelligibility when presented alone in the absence of the other modality

This same information can, when combined across modalities, provide good intelligibility (63% average accuracy)

When the audio signal leads the video, intelligibility falls off rapidly as a function of modality asynchrony

When the video signal leads the audio, intelligibility is maintained for asynchronies as long as 200 ms

There are many potential interpretations of the data

The simplest explanation is also the most ecologically valid: we prefer conditions where visual information leads audio information because that's what we're confronted with in the real world. But there are still questions remaining.

For visual lag conditions, the similarity between within-modality and across-modality integration is intriguing, and may suggest an interpretation based on the linguistic information conveyed by A and V modalities and which modality first activates the system

That's All, Folks

Many Thanks for Your Time and Attention

VARIABILITY AMONG SUBJECTS

One Further Wrinkle to the Story ....

Perhaps the most intriguing property of the experimental results concerns the intelligibility patterns associated with individual subjects

One Further Wrinkle to the Story ....

Perhaps the most intriguing property of the experimental results concerns the intelligibility patterns associated with individual subjects

For eight of the nine subjects, the condition associated with the highest intelligibility was one in which the video signal led the audio

One Further Wrinkle to the Story ….

Perhaps the most intriguing property of the experimental results concerns the intelligibility patterns associated with individual subjects

For eight of the nine subjects, the condition associated with the highest intelligibility was one in which the video signal led the audio

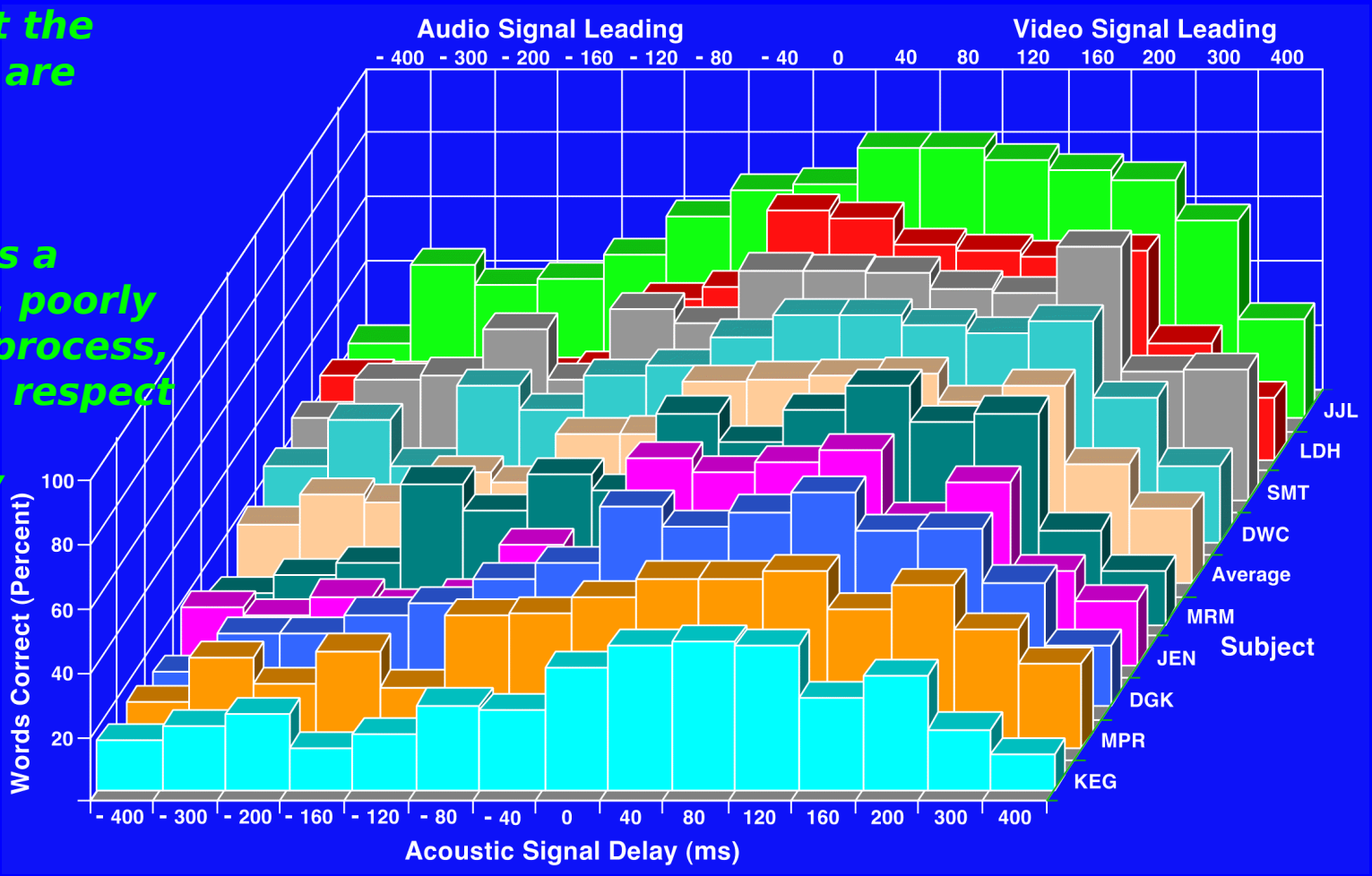
The length of optimal asynchrony (in terms of intelligibility) varies from subject to subject, but is generally between 80 and 120 ms

Auditory-Visual Integration - by Individual S's

These data are complex, but the implications are clear.

Audio-visual integration is a complicated, poorly understood process, at least with respect to speech intelligibility

Video signal leading is better than synchronous for 8 of 9 subjects



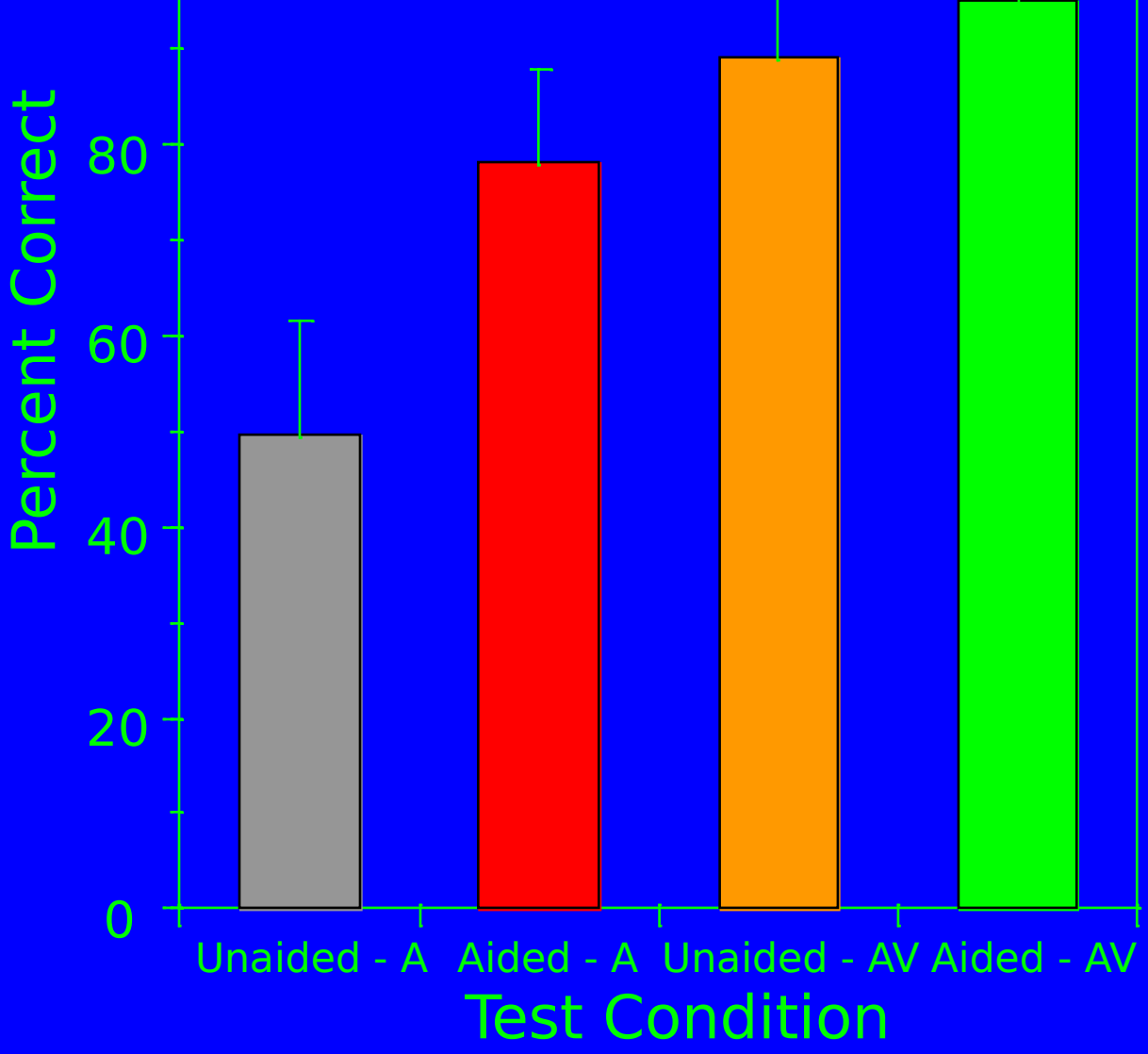
Variation across subjects

Hearing Aids and Speechreading

Are hearing aids optimally designed and fit for auditory-visual speech input?

- *How can they be when visual cues are not considered in the design or fitting process!*

Consonant Recognition - Amplification versus Speechread



Walden, Grant, & Cord (2001). Ear and Hearing **22**, 333-341.

Hearing Aid Summary

To optimize AV speech recognition, hearing aids must...

- Provide accurate reception of speech cues not readily available through speechreading *(complementary cues)*
  - voicing
  - manner
- There may have to be a trade off between providing the best auditory alone experience and one where AV cues are optimized